Species Distribution Modelling of Corals and Sponges from Research Vessel Survey Data in the Newfoundland and Labrador Region for Use in the Identification of Significant Benthic Areas

J. Guijarro, L. Beazley, C. Lirette, E. Kenchington, V. Wareham, K. Gilkinson, M. Koen-Alonso, F.J. Murillo

Ocean and Ecosystem Sciences Division
Maritimes Region
Fisheries and Oceans Canada

Bedford Institute of Oceanography
PO Box 1006
Dartmouth, Nova Scotia
Canada B2Y 4A2

2016

Canadian Technical Report of Fisheries and Aquatic Sciences 3171

Fisheries and Oceans Canada | Pêches et Océans Canada

Canada

Canadian Technical Report of Fisheries and Aquatic Sciences

Technical reports contain scientific and technical information that contributes to existing knowledge but which is not normally appropriate for primary literature. Technical reports are directed primarily toward a worldwide audience and have an international distribution. No restriction is placed on subject matter and the series reflects the broad interests and policies of Fisheries and Oceans Canada, namely, fisheries and aquatic sciences.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is abstracted in the data base *Aquatic Sciences and Fisheries Abstracts*.

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page.

Numbers 1-456 in this series were issued as Technical Reports of the Fisheries Research Board of Canada. Numbers 457-714 were issued as Department of the Environment, Fisheries and Marine Service, Research and Development Directorate Technical Reports. Numbers 715-924 were issued as Department of Fisheries and Environment, Fisheries and Marine Service Technical Reports. The current series name was changed with report number 925.

Rapport technique canadien des sciences halieutiques et aquatiques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles, mais qui ne sont pas normalement appropriés pour la publication dans un journal scientifique. Les rapports techniques sont destinés essentiellement à un public international et ils sont distribués à cet échelon. Il n'y a aucune restriction quant au sujet; de fait, la série reflète la vaste gamme des intérêts et des politiques de Pêches et Océans Canada, c'est-à-dire les sciences halieutiques et aquatiques.

Les rapports techniques peuvent être cités comme des publications à part entière. Le titre exact figure au-dessus du résumé de chaque rapport. Les rapports techniques sont résumés dans la base de données *Résumés des sciences aquatiques et halieutiques*.

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement auteur dont le nom figure sur la couverture et la page du titre.

Les numéros 1 à 456 de cette série ont été publiés à titre de Rapports techniques de l'Office des recherches sur les pêcheries du Canada. Les numéros 457 à 714 sont parus à titre de Rapports techniques de la Direction générale de la recherche et du développement, Service des pêches et de la mer, ministère de l'Environnement. Les numéros 715 à 924 ont été publiés à titre de Rapports techniques du Service des pêches et de la mer, ministère des Pêches et de l'Environnement. Le nom actuel de la série a été établi lors de la parution du numéro 925.

Canadian Technical Report of
Fisheries and Aquatic Sciences 3171

2016

Species Distribution Modelling of Corals and Sponges from Research Vessel Survey Data in the Newfoundland and Labrador Region for Use in the Identification of Significant Benthic Areas

by

J. Guijarro[1], L. Beazley[1], C. Lirette[1], E. Kenchington[1], V. Wareham[2], K. Gilkinson[2], M. Koen-Alonso[2], F.J. Murillo[1]

Fisheries and Oceans Canada

[1]Ocean and Ecosystem Sciences Division
Maritimes Region
Bedford Institute of Oceanography
P.O. Box 1006
Dartmouth, Nova Scotia
B2Y 4A2

[2]Northwest Atlantic Fisheries Centre
Fisheries and Oceans Canada
80 East White Hills Road, PO Box 5667
St. John's, Newfoundland and Labrador
A1C 5X1

© Her Majesty the Queen in Right of Canada, 2016.
Cat. No. Fs97-6/3171E-PDF ISBN 978-0-660-05651-7 ISSN 1488-5379

Correct citation for this publication:

Guijarro, J., Beazley, L., Lirette, C., Kenchington, E., Wareham, V., Gilkinson, K., Koen-Alonso, M., and Murillo, F.J. 2016. Species Distribution Modelling of Corals and Sponges from Research Vessel Survey Data in the Newfoundland and Labrador Region for Use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3171: vi + 126p.

TABLE OF CONTENTS

ABSTRACT v
RÉSUMÉ vi
INTRODUCTION 1
MATERIAL AND METHODS 3
Study Area 3
Environmental Data Layers 4
Response Data 5
Random Forest Modelling 10
Model Evaluation 11
Model Extrapolation 13
Ecological Interpretation 13
Alternative Prediction Models 14
RESULTS 14
Sponges (Porifera) 14
Data Sources and Distribution 14
Model 1 – Balanced Species Prevalence 16
Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence 24
Model Selection 30
Prediction of Sponge Biomass Using Random Forest 30
Sea Pens (Pennatulacea) 35
Data Sources and Distribution 35
Model 1 – Balanced Species Prevalence 37
Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence 44
Model Selection 50
Prediction of Sea Pen Biomass Using Random Forest 50
Large Gorgonian Corals 55
Data Sources and Distribution 55
Model 1 – Balanced Species Prevalence 57
Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence 64
Model Selection 70
Prediction of Large Gorgonian Coral Biomass Using Random Forest 70
Small Gorgonian Corals 75
Data Sources and Distribution 75
Model 1 – Balanced Species Prevalence 77
Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence 83
Model Selection 89
Prediction of Small Gorgonian Coral Biomass Using Random Forest 89
DISCUSSION 94
ACKNOWLEDGMENTS 98
REFERENCES 98
APPENDIX 1 103
Alternative Prediction Models- Generalized Additive Models for Predicting Coral and Sponge Biomass in the Newfoundland and Labrador Region 103

Sponges (Porifera) 104
Sea Pens (Pennatulacea) 107
Large Gorgonian Corals 111
Small Gorgonian Corals 116
APPENDIX 2 121
Congruence between Fisheries Observer Data and Species Presence Probability 121

ABSTRACT

Guijarro, J., Beazley, L., Lirette, C., Kenchington, E., Wareham, V., Gilkinson, K., Koen-Alonso, M., and Murillo, F.J. 2016. Species Distribution Modelling of Corals and Sponges from Research Vessel Survey Data in the Newfoundland and Labrador Region for Use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3171: vi + 126p.

We used a species distribution modelling approach called random forest (RF) to predict the probability of occurrence and biomass of sponges, sea pens, and large and small gorgonian corals across the entire spatial extent of Fisheries and Oceans, Canada's (DFO) Newfoundland and Labrador Region. A suite of 66 environmental variables from different data sources were used. Models utilized catch records from the DFO multispecies trawl survey, DFO/industry northern shrimp surveys, and Spanish trawl surveys. Most models had excellent predictive capacity with cross-validated Area Under the Receiver Operating Characteristic Curve (AUC) values ranging from 0.786 to 0.926. Areas of suitable habitat were identified for each taxon and were contrasted against their known distribution. Generalized additive models (GAMs) were developed to predict the biomass distribution of each taxonomic group and serve as a comparison to the RF models. The RF and GAM models provided similar results, although GAMs provided superior predictions of biomass along the slopes of Newfoundland and Labrador for some taxonomic groups. Aside from providing continuous prediction maps of significant benthic taxa for the entire Newfoundland and Labrador Region that will be useful in ecosystem management decision-making processes, these results could be used to refine the outer boundaries of significant concentrations of these organisms identified by kernel density analyses and identify new suitable habitat not sampled by the trawl surveys.

RÉSUMÉ

Guijarro, J., Beazley, L., Lirette, C., Kenchington, E., Wareham, V., Gilkinson, K., Koen-Alonso, M., et Murillo, F.J. 2016. Modélisation de la répartition des espèces de coraux et d'éponges à partir des données des relevés par navire de recherche dans la région de Terre-Neuve-et-Labrador aux fins d'utilisation dans la détermination des zones benthiques importantes. Can. Tech. Rep. Fish. Aquat. Sci. 3171: vi + 126p.

Nous nous sommes servis d'une approche de modélisation de la répartition des espèces appelée le modèle de forêts aléatoires pour prévoir la probabilité de la présence et de la répartition de la biomasse des éponges, des pennatules et des bancs de grandes et petites gorgones dans l'ensemble de l'étendue spatiale de Pêches et Océans Canada (MPO), région de Terre-Neuve et du Labrador. Un ensemble de 66 variables environnementales de différentes sources de données a été utilisé. Les modèles utilisent des registres de pêches tirés des relevés plurispécifiques au chalut de Pêches et Océans Canada (MPO), des relevés sur la crevette nordique et sur le chalutier espagnol du MPO et de l'industrie. La plupart des modèles avaient une excellente efficacité de prévision selon des valeurs contre-validées de l'aire sous la courbe de la fonction d'efficacité du récepteur variant de 0,786 à 0,926. Les zones constituant un habitat convenable ont été déterminées pour chaque taxon et ont été mises en contraste par rapport à l'aire de répartition de l'espèce. Des modèles additifs généralisés ont été élaborés pour prédire la répartition de la biomasse de chaque groupe taxonomique et servent de points de comparaison aux modèles RF. Les résultats obtenus par les modèles RF et les modèles additifs généralisés étaient similaires, cependant les prévisions de la biomasse le long des talus de Terre-Neuve et du Labrador pour certains groupes taxonomiques par les modèles additifs généralisés étaient meilleures. En plus de fournir des cartes de prévision continue de taxons benthiques importantes pour l'ensemble de la région de Terre-Neuve et du Labrador qui seront utiles dans les processus décisionnels en matière de gestion de l'écosystème, ces résultats pourraient servir à préciser les limites extérieures des concentrations importantes de ces organismes déterminés par des analyses de noyaux de densité et à repérer d'autres habitats convenables qui n'avaient pas été échantillonnés par les relevés au chalut.

INTRODUCTION

In 2006, the United Nations General Assembly (UNGA) Resolution 61/105 on sustainable fisheries drew attention to the state of deep-water ecosystems and called upon member States and Regional Fisheries Management Organizations (RFMOs) to identify and protect vulnerable marine ecosystems (VMEs) in the high seas. To assist States and RFMOs in the implementation of Resolution 61/105, the Food and Agriculture Organization (FAO) of the United Nations provided examples of species groups, communities, and habitats considered to epitomize VMEs in the high seas (FAO, 2009). Among these are deep-water corals, sponge aggregations, and seep and hydrothermal vent communities. Through initiatives of the Northwest Atlantic Fisheries Organization (NAFO), the RFMO responsible for fisheries management in the high seas of the northwest Atlantic, considerable effort has been made towards mapping the distribution of deep-water coral and sponge VMEs in this area (e.g. Murillo et al., 2011; Barrio-Frojan et al., 2012; Murillo et al., 2012; Beazley et al., 2013; Beazley et al., 2015; Knudby et al., 2013a,b; NAFO, 2013; Kenchington, 2014; Kenchington et al., 2014). As a result, NAFO has closed 13 areas within its fishing footprint for the protection of VMEs until their review in 2020 (NAFO, 2015).

Canada is legally obligated to take action in response to UNGA Resolution 61/105 and other international agreements to identify and protect sensitive benthic marine species and habitats. In 2009, Fisheries and Oceans Canada (DFO) developed the Policy for Managing the Impacts of Fishing on Sensitive Benthic Areas. Guided by the *Oceans Act*, *Fisheries Act*, and other legislation for the management of fisheries and habitat resources, the Policy for Managing the Impacts of Fishing on Sensitive Benthic Areas called for the development of an Ecological Risk Analysis Framework to analyze the impacts of commercial, recreational, and Aboriginal fisheries on sensitive benthic habitat and species both within and outside Canada's 200 nautical mile Exclusive Economic Zone (EEZ). The policy outlined a two-step process for identifying Sensitive Benthic Areas: 1) determination of ecological or biological significance of the area, and 2) determination of the sensitivity of the area to proposed or ongoing fishing activity. The policy specifically highlights the need for improved knowledge on the location and type of benthic species, particularly in frontier areas, i.e., areas where no current fishing activity takes place and little or no available information on the benthic habitat, communities, or species. Due to the high level of uncertainty, frontier areas receive a higher level of risk aversion in order to reduce the potential impacts of fishing activities.

In 2013, DFO identified fifteen Ecologically or Biologically Significant Areas (EBSAs) in the Newfoundland and Labrador Shelves Bioregion (DFO, 2013), a Large Ocean Management Area (LOMA) that lies off the northeast coast of Newfoundland and the coast of Labrador. Three of these EBSAs are in coastal areas, seven in offshore areas on the outer banks and slope, four straddle both inshore and offshore regions, and one that is considered a transitory EBSA that follows the southern extent of pack ice. Additionally, eleven EBSAs were identified within the Placentia Bay-Grand Banks LOMA that lies south of the Newfoundland and Labrador Shelves Bioregion (Templeman, 2007). Several biological and oceanographic layers were considered to help evaluate and identify these EBSAs, however most were designated based on the aggregation of one or more important species (DFO, 2013). It was recognized that survey coverage in these areas was limited temporally and spatially and that the addition of new data could result in further EBSA designation and/or refinement to current EBSA boundaries, highlighting the need

for increased knowledge on the distribution of species for use in oceans planning and management processes.

Kenchington (2014) compiled information on marine benthic species and habitats occurring on the Scotian Shelf that are recognized in other jurisdictions as meeting EBSA or similar criteria. Many of these areas have the same or similar species groups as are found in Newfoundland and Labrador. Fourteen structure-forming, biogenic habitats were identified, including those formed by aggregations of sponges and deep-water corals. In the Newfoundland region, archiving of deep-water coral collections began in 2001 (Campbell and Simms, 2009). A core research program on deep-water corals led by DFO and Memorial University was significantly expanded in 2005 through funding from DFO's International Governance Strategy (IGS) Program. As a result, several works (e.g. Edinger et al., 2007; Hamoutene et al., 2007; Wareham and Edinger, 2007; Sherwood et al., 2008; Gilkinson and Edinger, 2009; Sherwood and Edinger, 2009; Hamel et al., 2010; Sun et al., 2010; Edinger et al., 2011; Mercier et al., 2011a,b; Sherwood et al., 2011; Baker et al., 2012a,b; Edinger and Sherwood, 2012) and a multi-disciplinary cruise that collected *in situ*, high-resolution remotely operated vehicle (ROV) camera and photo data along the continental slopes of Newfoundland in 2007 significantly improved our knowledge of the distribution of deep-water corals in the Newfoundland and Labrador Region. However, knowledge gaps still exist, particularly in the northern portion of the Newfoundland and Labrador Shelves Bioregion along the NAFO 2G-0B border (Wareham, 2009; Wareham et al., 2010) and in deeper waters beyond the continental shelf and slope (DFO, 2013).

Species distribution modelling (SDM) tools are becoming more widely considered in fisheries and habitat management processes for the identification of areas containing species and habitats of biological or ecological importance. SDMs can be used to predict the distribution of a species or group of species in unsampled areas based on their relationship with the environment in sampled areas. A number of different modelling approaches are currently available (see Guisan and Zimmerman (2000) for a review). Included in these are a series of non-parametric techniques, amongst which random forest (RF; Breiman, 2001) is considered one of the more superior methods (Cutler et al., 2007). SDMs using random forest have been recently applied in the northwest Atlantic to predict the distribution of sponge grounds as determined using a biomass threshold applied to research vessel trawl catch data (Knudby et al., 2013a). Unsampled areas along the Newfoundland and Labrador slopes were identified as having a moderate/high presence probability of sponge grounds. Random forest was also used to model black corals, large gorgonian corals, and sea pens within the NAFO Regulatory Area (see Knudby et al., 2013b). To our knowledge, SDM techniques have not been applied previously to deep-water corals within Canadian waters off Newfoundland and Labrador. Such models would serve to fill gaps in the distribution of these organisms and to complement other tools for the identification of significant concentrations of corals and sponges.

Here, we employed random forest models to predict the probability of occurrence and biomass of sponges, sea pens, and large and small gorgonian corals across the Newfoundland Marine Protected Area (MPA) Network Planning Area (herein referred to as the 'Newfoundland and Labrador Region'), which combines the spatial extent of both the Placentia Bay-Grand Bank and Newfoundland and Labrador Shelves LOMAs. Data from DFO research vessel multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish surveys were used. We

utilized an extensive suite of environmental predictor variables that were compiled specifically for the purposes of SDM in this region. With the exception of small gorgonian corals, these taxonomic groups are also considered benthic EBSA's as identified in Kenchington (2014), and all groups are considered VME indicators by NAFO (NAFO, 2013). Aside from providing continuous prediction maps for the entire Newfoundland and Labrador Region that will be useful in ecosystem management decision-making processes, the results in this report could be used to refine the outer boundaries of the significant concentrations as identified by kernel density estimation (Kenchington et al., 2016) and identify new areas that are not sampled by the trawl surveys.

MATERIAL AND METHODS

Study Area

The full spatial extent of DFO's Newfoundland and Labrador LOMA (termed the 'Newfoundland and Labrador (NL) Region' herein) was used as the boundary for species distribution modelling in this report (Figure 1). This extent is delimited by the 200 nautical mile EEZ in the east, and DFO's Maritimes Region and Central and Arctic administrative boundaries in the southwest and north, respectively. A 20-km buffer was placed around all land to avoid its inclusion in the models. The total area covered in the study extent is approximately 1,012,900 km^2 based on a NAD 1983 UTM Zone 21N projection.

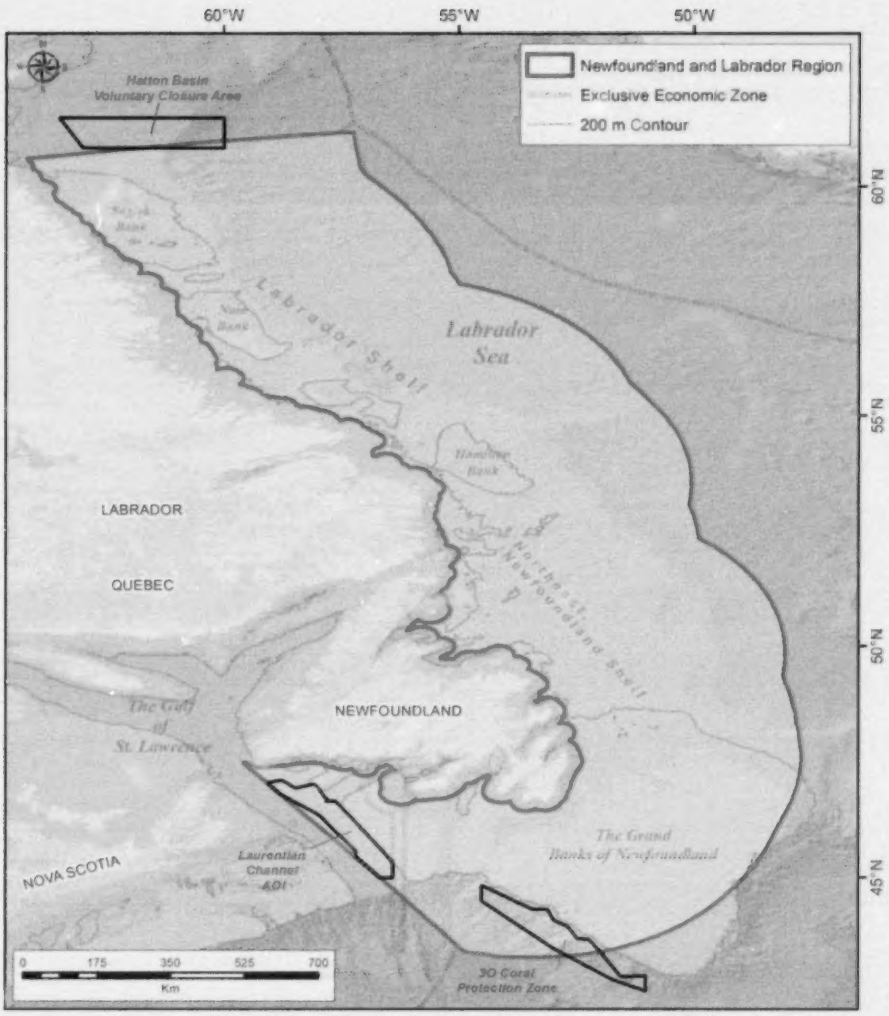

Figure 1. Extent of the boundary used for species distribution modelling (grey polygon) in the Newfoundland and Labrador Region. Place names are indicated on the map along with current closure areas within the Region.

Environmental Data Layers

Sixty-six environmental variables derived from various sources and native spatial resolutions were used as predictor variables in the random forest models (Table 1). Variables were chosen based on their availability and assumed relevance to the distribution of benthic fauna. Bathymetry was derived from the Canadian Hydrographic Service (CHS) Atlantic Bathymetry Compilation (ABC). This data is the highest resolution bathymetry available for the entire study area. In the Newfoundland and Labrador Region the data are resolved to 15 arc-seconds which is equivalent to approximately 500 m. Slope in degrees was derived from the depth raster using the 'Slope' tool in ArcMap's Spatial Analyst toolbox, ArcMap version 10.2.2 (ESRI, 2011). All other environmental variables were derived from long-term modelled oceanographic or remote-

sensing data and were spatially interpolated across the study area using ordinary kriging in ArcMap. Specific details on the methods used for the spatial interpolation of these variables are documented in a separate technical report (in prep., although see Beazley et al., 2016b for information on the same environmental data sources and variables for the Gulf of St. Lawrence). Only variables that were spatially interpolated with reasonable confidence were used in this report, and a number of variables (e.g., dissolved oxygen, silicate) were not considered. All predictor layers were displayed in raster format with geographic coordinates using the WGS 1984 datum and a ~0.015° cell size (approximately equal to 1 km horizontal resolution in the Newfoundland and Labrador Region).

Response Data

Species composition, as determined at sea, of the four taxonomic groups modelled in this report is presented in Table 2. These are presented for purposes of re-extracting the data and should not be considered as taxonomically certain. For each group, presence-absence records were derived from catch data from three different sources: 1) DFO research vessel multispecies trawl surveys conducted on the CCGS *Needler*, *Templeman* or *Teleost*, 2) DFO/industry Northern Shrimp Survey conducted on fishing vessels *Cape Ballard*, *Aqviq*, or *Kinguk*, and 3) Spanish research vessel groundfish trawl surveys conducted on the RV *Vizconde de Eza*. All tows were conducted following a stratified random design using Campelen trawl gear. DFO invertebrate catch data were provided by DFO's Newfoundland and Labrador Region where they are archived and managed. Spanish groundfish catch data were provided by the Instituto Español de Oceanografía (IEO) based in Vigo, Spain. Data were available from 1995 to 2015 for the sponges, and from 2003 to 2015 for all other taxonomic groups. Absence records were created from null (zero) catches that occurred in the same surveys.

Table 1. Summary of the 66 environmental variables used as predictor variables in random forest modelling. N/A = Not Applicable.

Variable	Data source	Temporal range	Unit	Native resolution
Depth	CHS-ABC	N/A	metres	15 arc-sec (~500 m)
Slope	CHS-ABC	N/A	degrees	15 arc-sec (~500 m)
Bottom Salinity Mean	GLORYS2V1	1993 - 2011	N/A	¼ °
Bottom Salinity Average Minimum	GLORYS2V1	1993 - 2011	N/A	¼ °
Bottom Salinity Average Maximum	GLORYS2V1	1993 - 2011	N/A	¼ °
Bottom Salinity Average Range	GLORYS2V1	1993 - 2011	N/A	¼ °
Bottom Temperature Mean	GLORYS2V1	1993 - 2011	°C	¼ °
Bottom Temperature Average Minimum	GLORYS2V1	1993 - 2011	°C	¼ °
Bottom Temperature Average Maximum	GLORYS2V1	1993 - 2011	°C	¼ °
Bottom Temperature Average Range	GLORYS2V1	1993 - 2011	°C	¼ °
Bottom Current Speed Mean	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Bottom Current Speed Average Minimum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Bottom Current Speed Average Maximum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Bottom Current Speed Average Range	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Bottom Shear Mean	GLORYS2V1	1993 - 2011	Pa	¼ °
Bottom Shear Average Minimum	GLORYS2V1	1993 - 2011	Pa	¼ °
Bottom Shear Average Maximum	GLORYS2V1	1993 - 2011	Pa	¼ °
Bottom Shear Average Range	GLORYS2V1	1993 - 2011	Pa	¼ °
Surface Salinity Mean	GLORYS2V1	1993 - 2011	N/A	¼ °
Surface Salinity Average Minimum	GLORYS2V1	1993 - 2011	N/A	¼ °
Surface Salinity Average Maximum	GLORYS2V1	1993 - 2011	N/A	¼ °
Surface Salinity Average Range	GLORYS2V1	1993 - 2011	N/A	¼ °
Surface Temperature Mean	GLORYS2V1	1993 - 2011	°C	¼ °
Surface Temperature Average Minimum	GLORYS2V1	1993 - 2011	°C	¼ °
Surface Temperature Average Maximum	GLORYS2V1	1993 - 2011	°C	¼ °
Surface Temperature Average Range	GLORYS2V1	1993 - 2011	°C	¼ °

Surface Current Speed Mean	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Surface Current Speed Average Minimum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Surface Current Speed Average Maximum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Surface Current Speed Average Range	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Maximum Average Mixed Layer Depth Fall	GLORYS2V1	1993 - 2011	metres	¼ °
Maximum Average Mixed Layer Depth Winter	GLORYS2V1	1993 - 2011	metres	¼ °
Maximum Average Mixed Layer Depth Spring	GLORYS2V1	1993 - 2011	metres	¼ °
Maximum Average Mixed Layer Depth Summer	GLORYS2V1	1993 - 2011	metres	¼ °
Fall Chlorophyll *a* Mean	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Fall Chlorophyll *a* Minimum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Fall Chlorophyll *a* Maximum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Fall Chlorophyll *a* Range	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Spring Chlorophyll *a* Mean	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Spring Chlorophyll *a* Minimum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Spring Chlorophyll *a* Maximum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Spring Chlorophyll *a* Range	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Summer Chlorophyll *a* Mean	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Summer Chlorophyll *a* Minimum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Summer Chlorophyll *a* Maximum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Summer Chlorophyll *a* Range	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Annual Chlorophyll *a* Mean	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Annual Chlorophyll *a* Minimum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Annual Chlorophyll *a* Maximum	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Annual Chlorophyll *a* Range	SeaWiFS Level-3, NASA's OceanColor	2001 - 2010	$mg\ m^{-3}$	9 km
Fall Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 - 2010	$mg\ C\ m^{-2}\ day^{-1}$	9 km
Fall Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 - 2010	$mg\ C\ m^{-2}\ day^{-1}$	9 km
Fall Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 - 2010	$mg\ C\ m^{-2}\ day^{-1}$	9 km
Fall Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 - 2010	$mg\ C\ m^{-2}\ day^{-1}$	9 km

Spring Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 - 2010	mg C m^{-2} day^{-1}	9 km

Table 2. Species composition in each of the four taxonomic groups modelled using random forest. Also shown are the species/taxon codes associated with data entry of the DFO multispecies and northern shrimp survey records. *Indicates taxon listed in Spanish surveys.

Taxonomic Group	Species/Taxon	Taxon Code
Sponges	Porifera	1101
Sea Pens (Pennatulacea)	*Anthoptilum**	5117
	Anthoptilum grandiflorum	8937
	Distichoptilum gracile	8932
	Funiculina quandrangularis	8938
	Halipteris finmarchica	8936
	Pennatula aculeata	8934
	Pennatula cf. *aculeata*	8934
	Pennatula grandis	8935
	Pennatula cf. *grandis*	8935
	Pennatula phosphorea	8933
	Pennatula cf. *phosphorea*	8933
	Pennatula sp.	8954
	Pennatulacea	8901
	Sea pen sp.	8901
	Umbellula sp.	8972
Large Gorgonian Corals	*Acanthogorgia**	5073
	Acanthogorgia armata	8907
	Acanthogorgia cf. *armata*	8907
	*Keratoisis**	5070
	Keratoisis grayi	8906
	Paragorgia arborea	8903
	Paragorgia cf. *arborea*	8903
	Paramuricea sp.	8912
	*Paramuricea placomus**	8940/5114
	Paramuricea cf. *placomus*	8940
	Plexauridae	5054
	Parastenella atlantica	8944
	Primnoa resedaeformis	8902
Small Gorgonian Corals	*Acanella arbuscula*	8909
	Anthothela grandiflora	8915
	Chrysogorgia cf. *agassizii*	8924
	Chrysogorgia sp.	8965
	Radicipes gracilis	8910
	Swiftia sp.	8959

The presence–absence records used in each random forest model (see below) were filtered so that only one presence or absence occurred within a single environmental data raster cell (~1 km). Presence records took precedence over an absence record when both occurred within the same raster cell.

Biomass (kg) data associated with the DFO and Spanish survey records were also extracted for use in regression random forest models. For each taxonomic group, weights were averaged across multiple tows occurring within the same environmental raster cell.

Random Forest Modelling

Random forest (Breiman, 2001), is a non-parametric machine learning technique, where multiple regression or classification trees (usually $\geq$ 500) are built using random subsets of the data (Figure 2). Each tree is fit to a bootstrap sample of the biological observations (i.e. the 'in-bag' observations), and the best split at each node is selected based on a randomly-chosen subset of predictor variables. Regression trees are used for response variables consisting of continuous data and classification trees for categorical variables. RF is a robust statistical method requiring no distributional assumptions on covariate relation to the response in comparison to other classical statistical models such as generalized linear models (GLM) or generalized additive models (GAM).

For classification with presence–absence response data, random forest can be used to predict the probability of a species' presence in non-sampled areas by identifying areas with similar environmental conditions. For regression with biomass response data, random forest can be used to predict the species' biomass in non-sampled areas by identifying areas with similar environmental conditions. The models were built in the statistical computing software package R (R Core Team, 2015) using the 'randomForest' package (Liaw and Wiener, 2002). Default values were used for RF parameters, using 500 trees.

The catch records for some taxonomic groups are characterized by a higher number of absences relative to presences (i.e., unbalanced species prevalence). The distribution of these two classes may be biased spatially and/or environmentally across the study area. Classification accuracy in random forest is prone to bias when the categorical response variable is highly imbalanced (Chen et al., 2004). This is due to over-representation of the majority class in the bootstrap sample leading to a higher frequency in which the majority class is drawn, therefore skewing predictions in that favour (Evans et al., 2011). Several different approaches have been used to address imbalanced data: 1) assign a high cost to misclassification of the minority class, 2) down-sample the majority class, and 3) up-sample the minority class (Evans et al., 2011). Although a number of studies suggest a balanced modelling prevalence of 0.5 (McPherson et al., 2004; Liu et al., 2005), this approach may result in a loss of information particularly for rare species, and may not be necessary when the model training data is reliable and not biased spatially and/or environmentally (Jiménez-Valverde and Lobo, 2006). Another widely-used approach is to adjust the threshold used to divide the probabilistic predictions of occurrence into discrete predictions of presence or absence, to match modelling prevalence (Liu et al., 2005). The latter approach has shown to produce constant error rates and optimal model accuracy measures compared to balancing modeling prevalence (Liu et al., 2005; Hanberry and He, 2013).

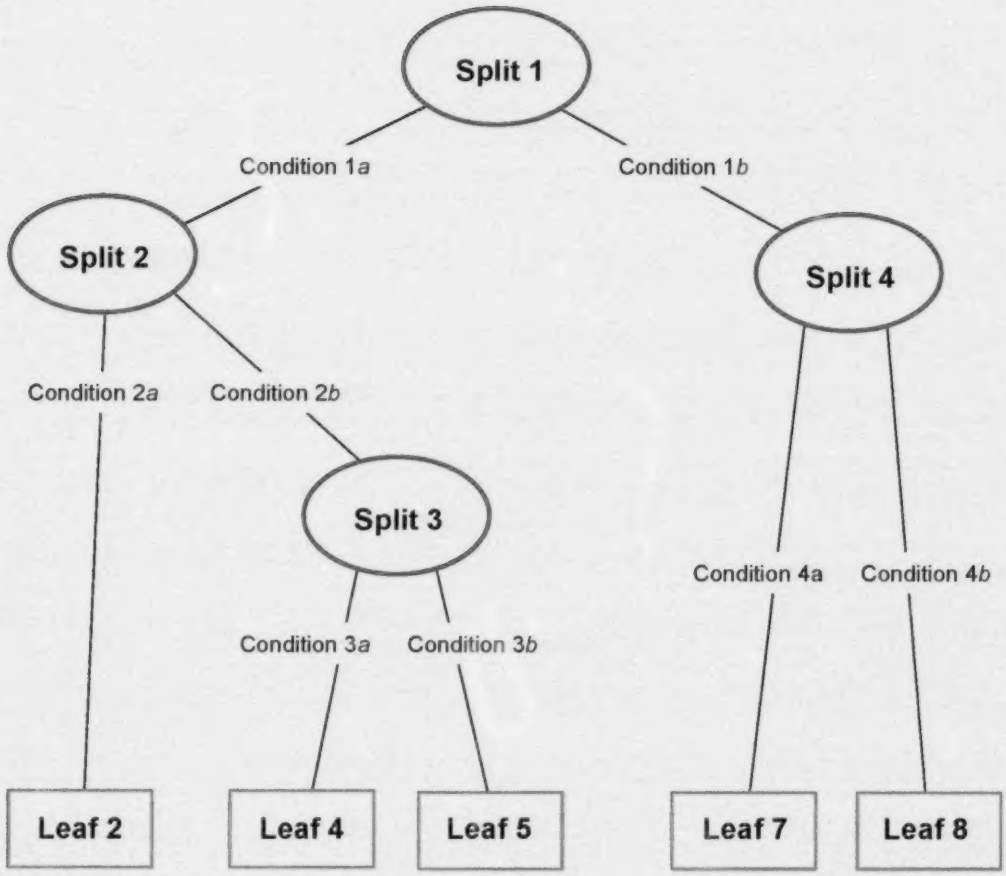

Figure 2. An example of a regression model tree (adapted from Kuhn and Johnson, 2013).

For each taxonomic group we assessed the number of presences and absences and their spatial distribution across the study area. We employed two different modelling methods. The first method was to model the response data with a balanced species prevalence and threshold of 0.5 (Model 1). Here the absence records were randomly down-sampled to match the number of presences prior to modelling. In the second method we used all presence and absence records and used species prevalence as the threshold (Model 2). The appropriateness of each modelling approach on the response data was assessed based on the model accuracy measures (see explanation below of model accuracy measures) and the spatial pattern of the predictions of presence probability in relation to the response data.

Model Evaluation

Presence-Absence Response Data – Classification Model

Accuracy measures were derived from validated data using 10-fold cross validation (10 resamples over which performance estimates were obtained). In 10-fold cross validation the

response data are randomly split into 10 equal-sized groups and the model is trained on a combination of 9, while validated on the remaining group.

Three measures of accuracy were used to assess model performance: 1) sensitivity, 2) specificity, and 3) AUC, or Area Under the Receiver Operating Curve. In a classification model with two classes (e.g. presence and absence), there are four possible predicted outcomes: 1) true positive, where observed presences are predicted as presences, 2) false negative, where observed presences are predicted as absences, 3) true negative, where observed absences are predicted as absences, and 4) false positive, where observed absences are predicted as presences (Fawcett, 2006). Sensitivity measures the proportion of observed presences correctly predicted as presence (i.e. the true positive rate) (McPherson et al., 2004; Fawcett, 2006). Low sensitivity indicates high omission error (i.e. false negative rate). Specificity measures the proportion of observed absences correctly predicted as absence (i.e. the true negative rate). Low specificity indicates high commission error (i.e. the false positive rate). Both sensitivity and specificity are derived from a two-by-two confusion matrix of the tabulated predicted outcomes.

The AUC is a threshold-independent measure of model accuracy that is calculated from the combination of true positive rate (sensitivity) and false positive rate (1 – specificity), and equals the probability that the model will rank a randomly-chosen presence instance higher than a randomly-chosen absence instance (Fawcett, 2006). Its value ranges from 0 to 1, with values larger than 0.5 indicating performance better than random (Fawcett, 2006). It was calculated using 10-fold cross validation.

For models generated using a balanced species prevalence and threshold of 0.5, 10 data subsets were created with same number of presence and absences (balanced data) and 10 models were run. AUC was determined by averaging AUC values between folds within each run. The model with the highest average AUC was considered the most accurate in predicting the validated data and was used as the final model in which predicted presence probabilities of the response data were generated. The predicted outcomes from the two-by-two confusion matrices were summed across all 10 folds to give a complete confusion matrix for each model from which sensitivity and specificity were calculated (resulting in 10 confusion matrices, one for each data subset). For models generated on unbalanced data but with a threshold equal to species prevalence, only one model was considered and the AUC was determined from each of the 10 validated datasets by averaging all the AUC values between folds. The predicted outcomes from the two-by-two confusion matrices were summed across all 10 model folds to give one confusion matrix from which sensitivity and specificity were calculated.

Biomass Response Data – Regression Model

Models were validated using 10-fold cross validation, whilst maintaining the range of biomass values in each split. Data were split using the createFolds function in R. This function performs stratified partitioning into k groups in order to evenly distributed the biomass within splits. Models were built using each calibrated and validated dataset and accuracy measures were calculated for each corresponding dataset. The accuracy measures used to validate the models included the goodness-of-fit statistic R^2, the root-mean-square error (RMSE) value and the percentage of variance explained. RMSE was normalized to a percentage of the range of observed biomass values ($y_{max} - y_{min}$) for each specific response (NRMSE) to facilitate the

comparison between responses in the different models. Cross validation gives an average of the accuracy measures used, but can also be used to estimate the variability around the mean to evaluate the stability of the model fit, and to check for the arbitrary effects from subsampling data for calibrate and validate the model.

Model Extrapolation

The spatial extent of the Newfoundland and Labrador Region reaches far beyond the continental shelf and slope, down to ~4360 m depth. Our data observations are limited to depths above ~1600 m. Extrapolation of model predictions to areas outside of the range of data observations may produce unreliable predictions in those areas (Elith et al., 2010). Random forest models average the decision across regression trees to predict piecewise constant functions, giving a constant value for inputs falling under each leaf. When extrapolating outside the domain of the training data, where different physical conditions from those used to train the model likely exist, random forest models predict the same value as they would for the closest value in the tree for which they had training data (Breiman et al., 1984). For each random forest model, we highlight those areas within the study extent where model predictions are extrapolated. We define areas of extrapolation as those areas where at least one environmental variable has values above or below its sampled range.

Ecological Interpretation

Ecological interpretation of the models was aided by predictor variable importance measures and partial dependence plots generated from the final model. In classification random forest, variable importance is measured as the mean decrease in Gini value, otherwise known as Gini impurity. When the response data are split into two child nodes based on a randomly-chosen variable, the data in the two descendent nodes are more homogeneous than that of the parent node. This difference in homogeneity between parent and child nodes is measured by the Gini index, where the increase in homogeneity equals a decrease in Gini value. The sum of all decreases in Gini index for each variable in each tree is averaged across all trees in the model 'forest' and then across all 10 repetitions of each model fold. The variable with the highest mean decrease in Gini value, or in other words, the variable that was used to split the data at the highest number of nodes, is considered the most important variable in the model.

Variable importance in regression random forest is measured by the mean decrease in the residual sum of squares when the variable is included in a tree split. Partial dependence plots using the partialPlot function in R were generated for the 6 highest variable importance scores. Partial dependence plots show the relationship between a particular predictor variable and log-transformed predicted probabilities of presence (for classification models) or the biomass regression function (for regression models),while the other predictor variables were held constant at their mean observed value and are useful in showing general trends in model accuracy's dependence on the predictors (Herrick et al., 2013). For classification models, the y axis ranges from $-\infty$ to ∞ and quantifies the log-odds of a positive classification for the total range of values in x. Log-odds are logarithmic transformations of the probabilities for values in x (Hastie et al., 2005). These values were transformed to the original presence probability scale using:

$$p = \exp(y) / (1 + \exp(y))$$

where p = the probability of presence, and y is the log-odds of presence, the standard output from the partialPlot function.

Alternative Prediction Models

Generalized additive models (GAMs) were developed to predict the biomass distribution of each taxonomic group. GAMs were developed to compare to the random forest results and to determine whether predictions could be improved for the areas considered as extrapolated by random forest models. Methodology and results for the GAM models are presented in Appendix 1.

RESULTS

Sponges (Porifera)

Data Sources and Distribution

Sponge catch data was collected over a span of 21 years from 1995 to 2015 (Table 3) and consisted of 3860 presence and 10,980 absence records. Absence records were distributed relatively evenly across the study extent, while presence records were highly concentrated on Hamilton Bank and the Northeast Newfoundland Shelf and along the slopes of Newfoundland and Labrador (Figure 3). Few presence records were distributed on Grand Bank and Saglek Bank off northern Labrador. The highest mean biomass records (up to 1226.29 kg) occurred east of Hamilton Bank. Several, smaller mean catches occurred on the Labrador Slope off Saglek Bank.

Table 3. Number of presence and absence records of sponge catch recorded from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 1995 and 2015.

Year	Total number of presences	Total number of absences
1995	45	149
1996	126	624
1997	145	426
1998	117	658
1999	83	584
2000	62	570
2001	45	637
2002	30	363
2003	60	369
2004	77	602
2005	114	714
2006	200	689
2007	206	612
2008	299	521
2009	348	548
2010	383	578
2011	310	586
2012	379	567
2013	391	553
2014	355	384
2015	85	246

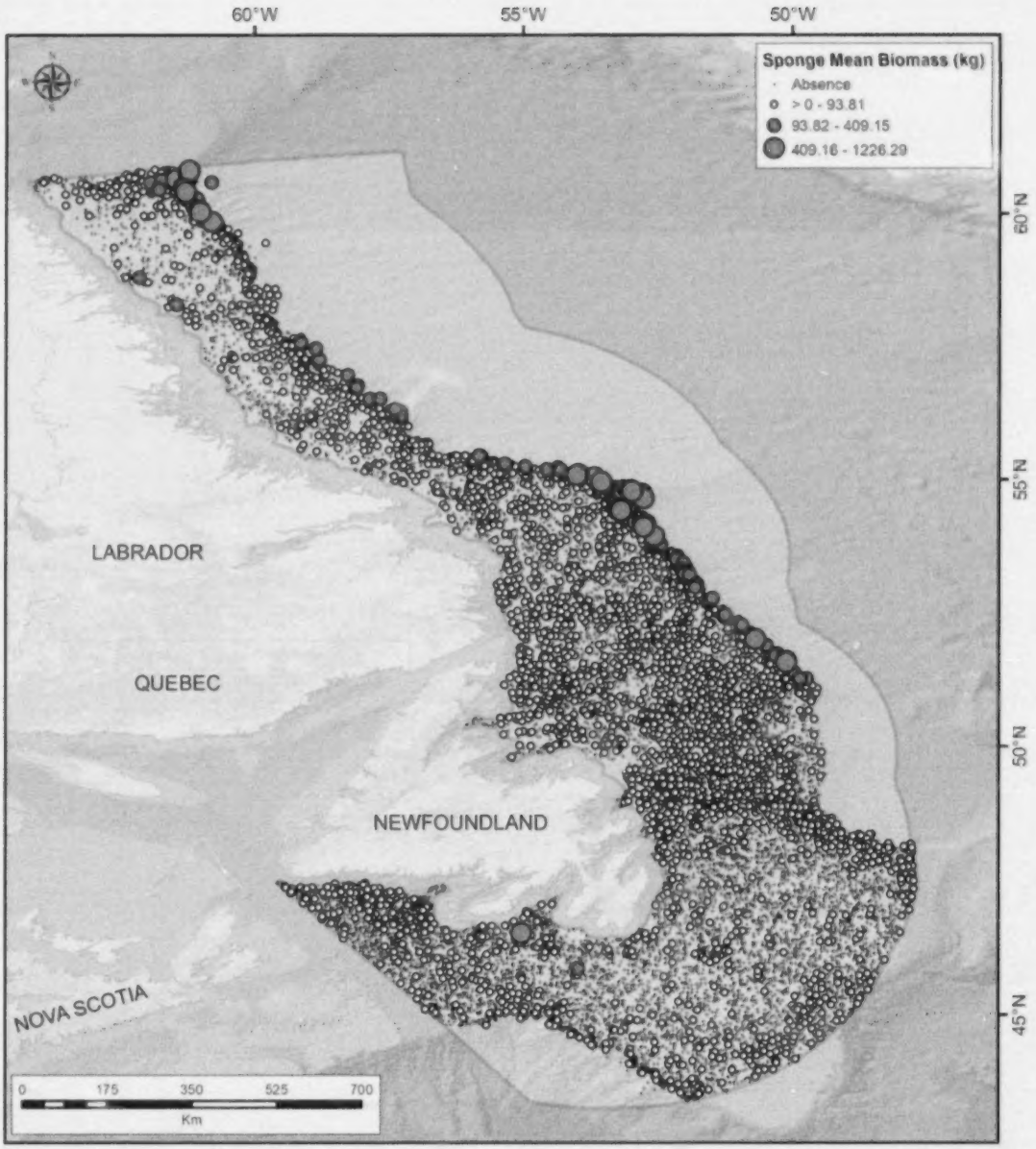

Figure 3. Mean biomass (kg) per grid cell of sponge data recorded from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 1995 and 2015. Also shown are absence records from the same surveys.

Model 1 – Balanced Species Prevalence

Model accuracy measures (mean AUC, sensitivity, and specificity) for the random forest model on balanced species prevalence (3860 presences and 3860 absences; Model 1) are presented in Table 4. The average AUC was 0.788, indicating fair model performance. The highest AUC of 0.792 was associated with Model Run 10. The sensitivity and specificity measures were 0.735

and 0.704, respectively. The confusion matrix of this model is also presented in Table 4. Class errors for both the presence and absence classes were relatively moderate (0.266 and 0.296, respectively).

Table 4. Accuracy measures for all 10 model repetitions of 10-fold across validation of a random forest model of presence and absence of sponges within the Newfoundland and Labrador Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 10) which is considered the optimal model for predicting the presence probability of sponges.

Model Run	AUC	Sensitivity	Specificity
1	0.791	0.740	0.704
2	0.787	0.738	0.704
3	0.788	0.734	0.704
4	0.786	0.728	0.700
5	0.787	0.731	0.700
6	0.786	0.743	0.700
7	0.787	0.726	0.702
8	0.791	0.740	0.702
9	0.790	0.730	0.708
10	**0.792**	**0.735**	**0.704**
Mean	**0.788**	**0.734**	**0.703**
SD	**0.002**	**0.006**	**0.003**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	2718	1142	3860	0.296
Presence	1025	2835	3860	0.266

The presence probability prediction surface of sponges is presented in Figure 4. The highest predictions of presence probability occurred along on the Labrador Slope off Hamilton Bank and the Northeast Newfoundland Shelf and on the slope off Saglek Bank. Hamilton Bank and the Northeast Newfoundland Shelf had a moderate to high predicted presence probability of sponges, while Saglek and Nain Banks, and The Grand Banks of Newfoundland had low predictions of presence probability. Areas of high presence probability corresponded well with the spatial distribution of the presence records (Figure 5). However, the model appears to greatly extrapolate areas of presence probability beyond the location of presence observations, particularly in deeper waters off the Labrador Slope.

Figure 6 shows the actual presence and absence data observations (3860 presences and 3860 absences) used in the optimal run of Model 1. There appeared to be no additional spatial bias in the presence and absence records caused by random down-sampling of the absence data. Areas

of extrapolation are also shown in this figure. All deep water beyond the slope was considered extrapolated area. The area of high predicted presence probability of sponges off the Labrador Slope was considered extrapolated area.

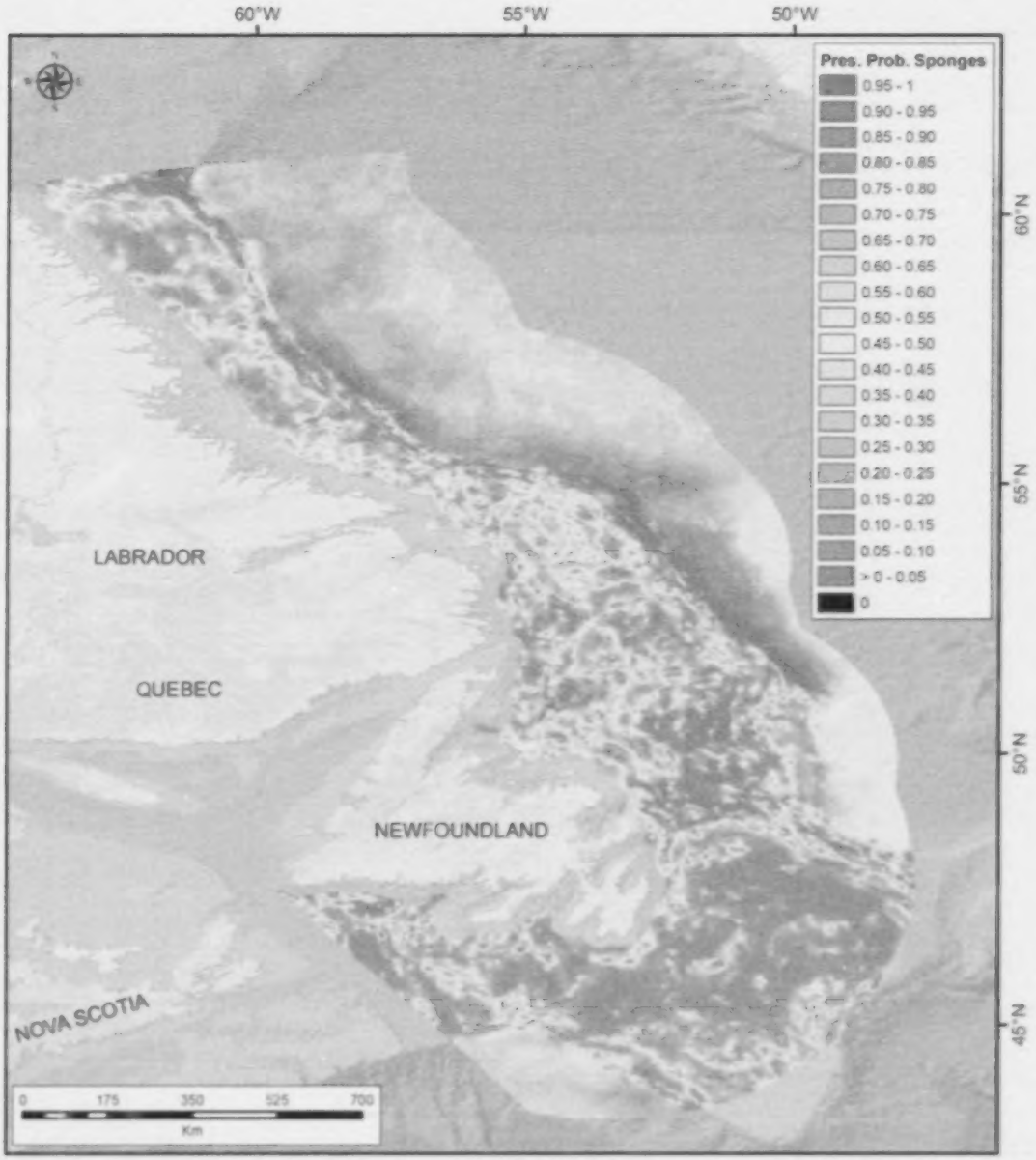

Figure 4. Predictions of presence probability from the optimal random forest model of sponge presence and absence data collected from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 1995 and 2015.

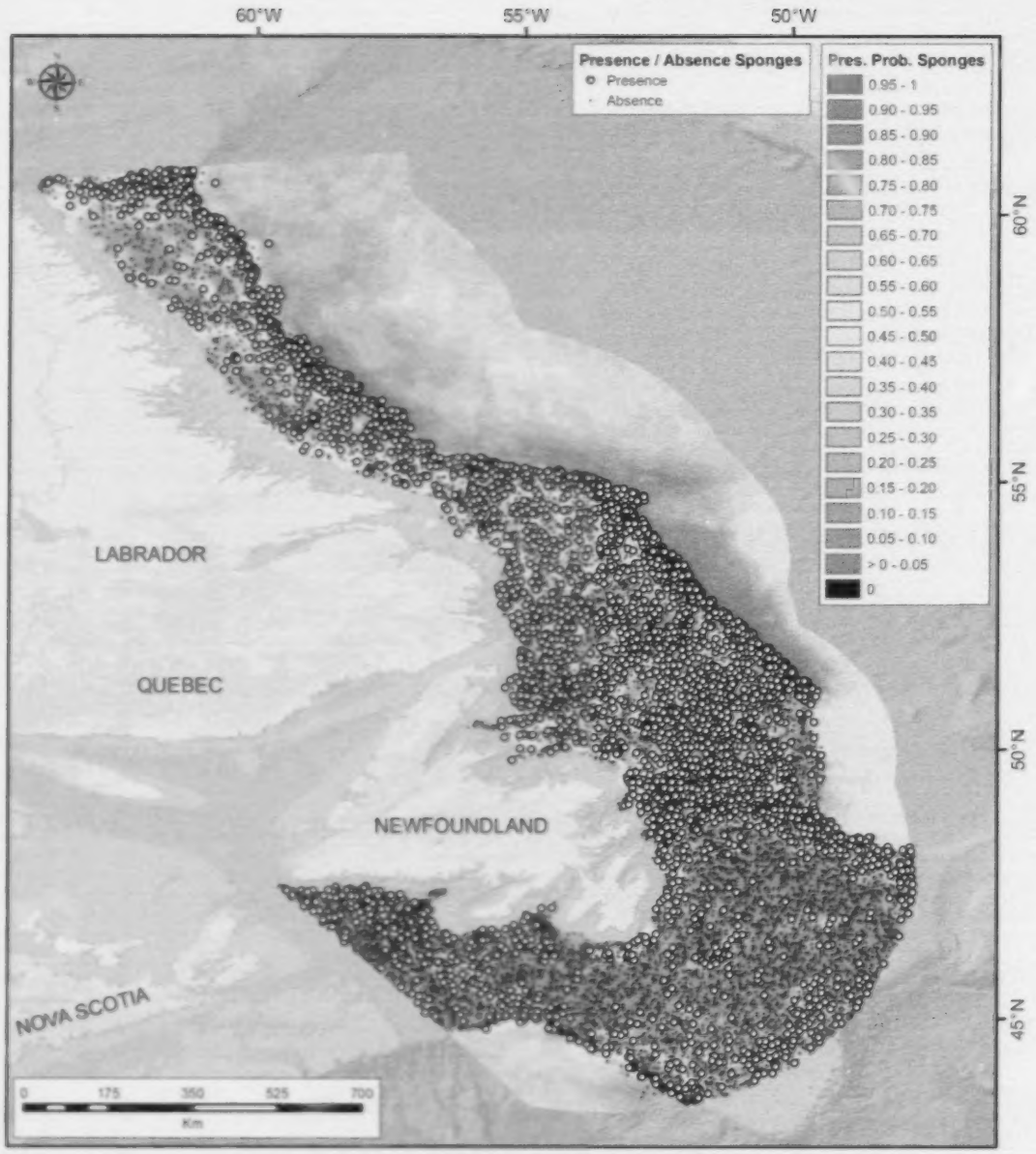

Figure 5. Presence and absence observations and predictions of presence probability of the optimal random forest model of sponge presence and absence data recorded from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 1995 and 2015.

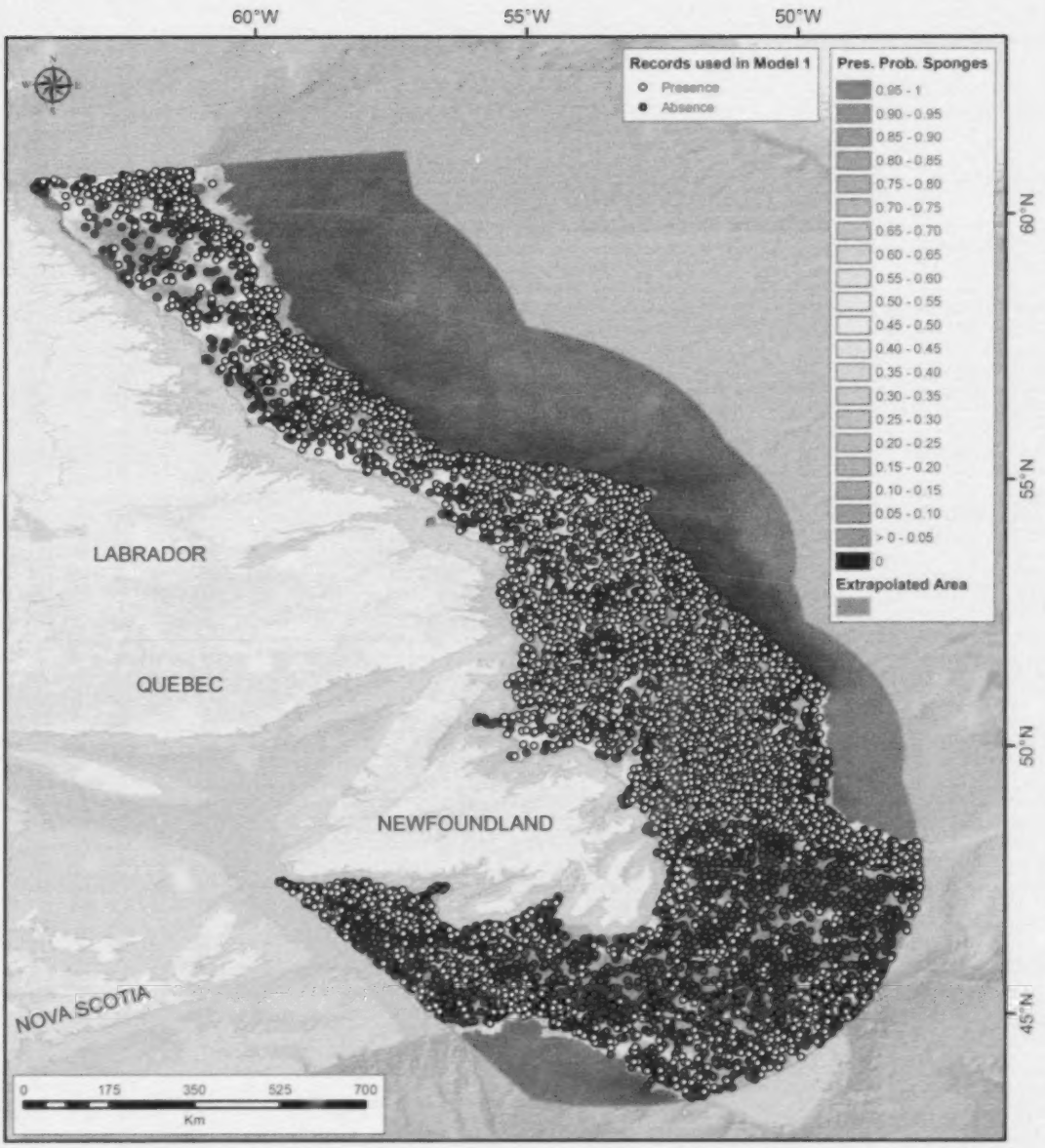

Figure 6. Map of the 7720 data observations (3860 presences and 3860 absences) of sponges used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (pres. prob.) of sponges generated from Model 1 and areas of model extrapolation.

Of all 66 environmental predictor variables used in the model, Fall Primary Production Average Maximum was the most important for the classification of the sponge presence and absence data (Figure 7). The distribution of this variable was relatively normal prior to spatial interpolation

(Guijarro et al., in prep). Examination of the Q-Q plot revealed no strong spatial pattern to the points over- and under-predicted by a normal distribution. Fall Primary Production Average Maximum was followed closely by Surface Temperature Average Maximum and Depth. Bottom salinity and temperature variables ranked high in the model. Partial dependence plots for the top 6 predictor variables are shown in Figure 8. The highest presence probability of sponges along the gradient in Fall Primary Production Average Maximum occurred between 300 and 600 mg C m^{-2} day^{-1}. Values in this range coincided with both over- and under-predicted values on the Labrador Shelf and Slope. These values are not of concern however, as there was a near-perfect fit between predicted and observed values in the kriging model, with only slight over-prediction of values between ~ 300 and 500 mg C m^{-2} day^{-1}. These over-predicted data points were still well within the range of high presence probability identified in the partial plot (Figure 8).

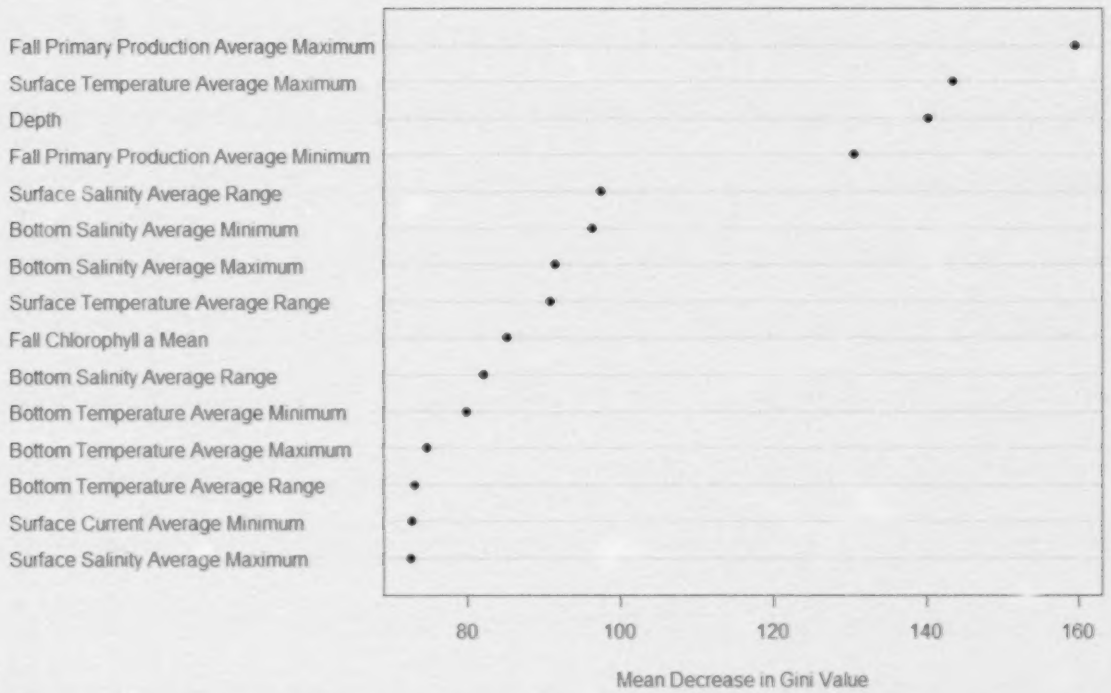

Figure 7. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting sponge presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini Value the more important the variable is for predicting the response data.

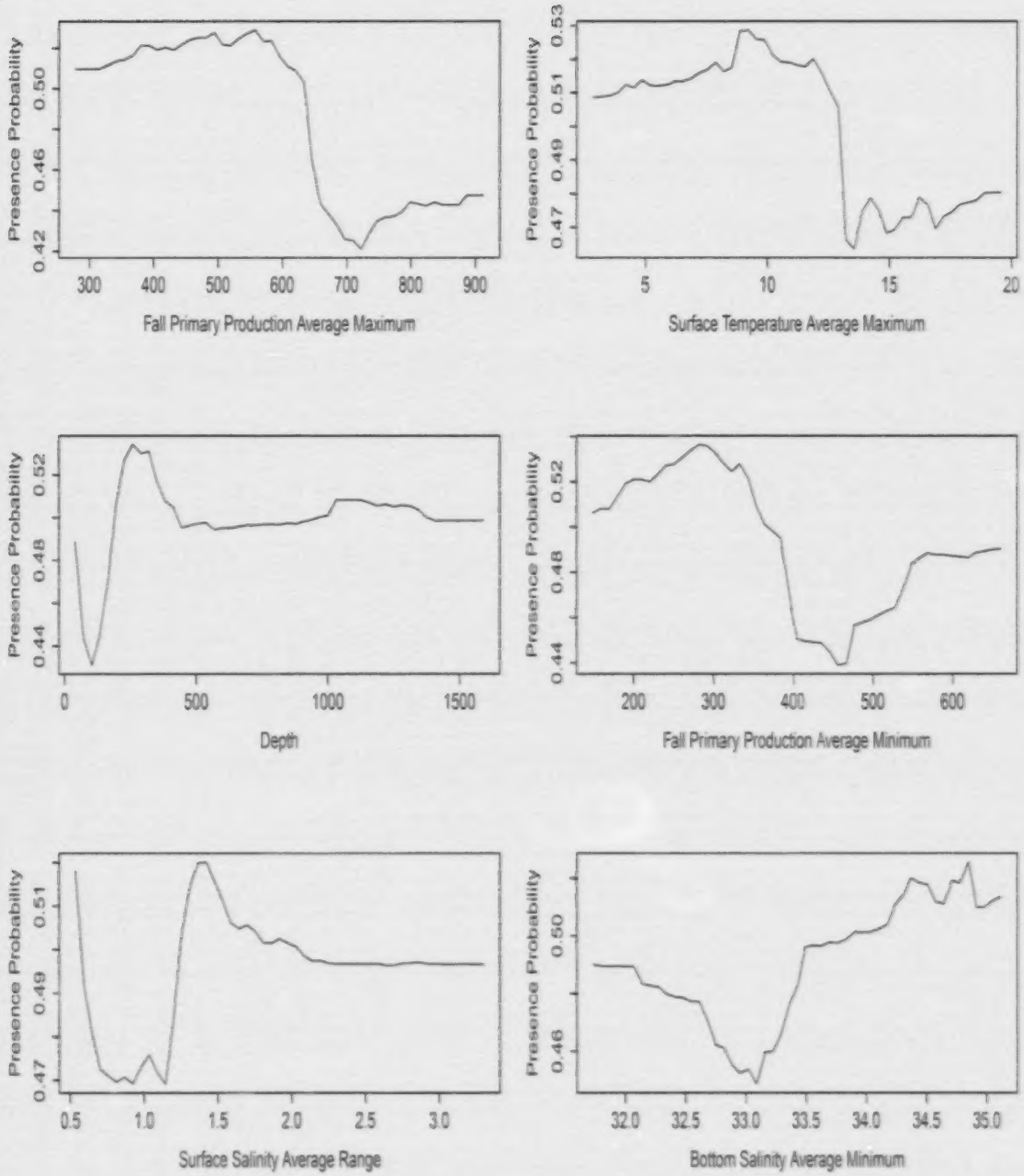

Figure 8. Partial dependence plots of the top 6 predictors from the optimal random forest model of sponge presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 5 shows the accuracy measures for the random forest model using all sponge presence and absence data (3860 presences 10,980 absences; Model 2) and a threshold equal to species prevalence (0.26). The average AUC calculated from this model was slightly lower than that of Model 1 (0.786 compared to 0.788 of Model 1). Sensitivity and specificity measures of Model 2 were also slightly lower than Model 1, and class error of the presence and absence classes was comparable to Model 1.

The predicted sponge presence probability surface generated from Model 2 is shown in Figure 9. The areas of high predicted presence probability from Model 1 are greatly reduced in this model. The highest sponge presence probabilities still occurred off the Labrador Slope and slope off Saglek Bank. However, the model does not appear to extrapolate high probabilities far beyond the location of presence observations (Figure 10), likely due to the inclusion of all absence records in the model. Figure 11 depicts the classification of sponge presence probability into presence and absence categories based on the prevalence threshold of 0.26. In this map, all presence probability values generated from Model 2 greater than 0.26 were classified as presence, while values less than 0.26 were classed as absence. With the exception of Nain and Saglek Banks, most of the shelf and slopes off Labrador were classified as sponge presence, while most of The Grand Banks of Newfoundland was classified as sponge absence.

Table 5. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model presence and absence of sponges within the Newfoundland and Labrador Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.773		**Absence**	**Presence**				
2	0.782	**Absence**	7728	3252	10980	0.296	0.729	0.704
3	0.792	**Presence**	1045	2815	3860	0.271		
4	0.787							
5	0.791							
6	0.776							
7	0.788							
8	0.803							
9	0.776							
10	0.797							
Mean	**0.786**							
SD	**0.010**							

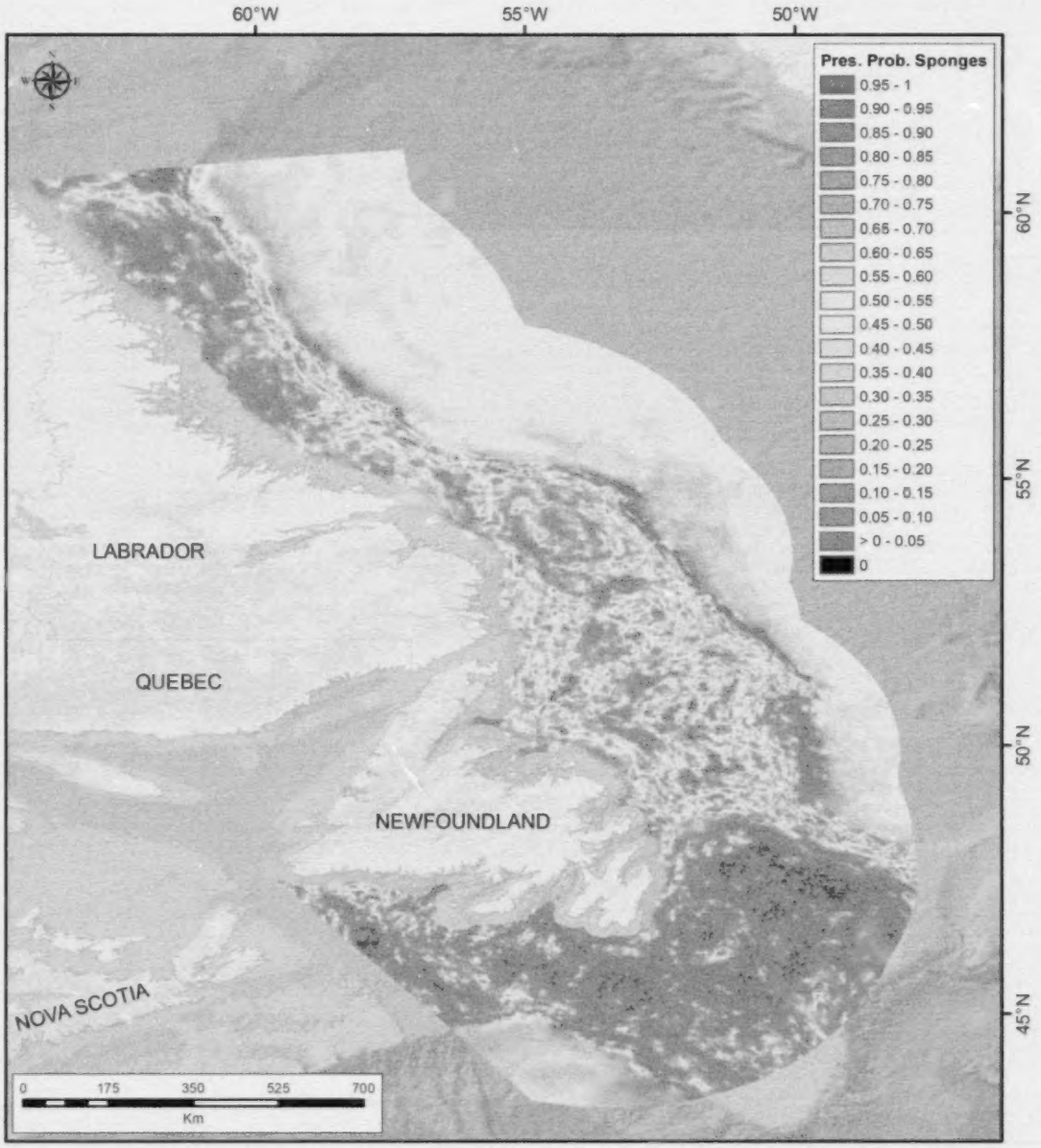

Figure 9. Predictions of presence probability from the unbalanced random forest model of sponge presence and absence data collected from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 1995 and 2015.

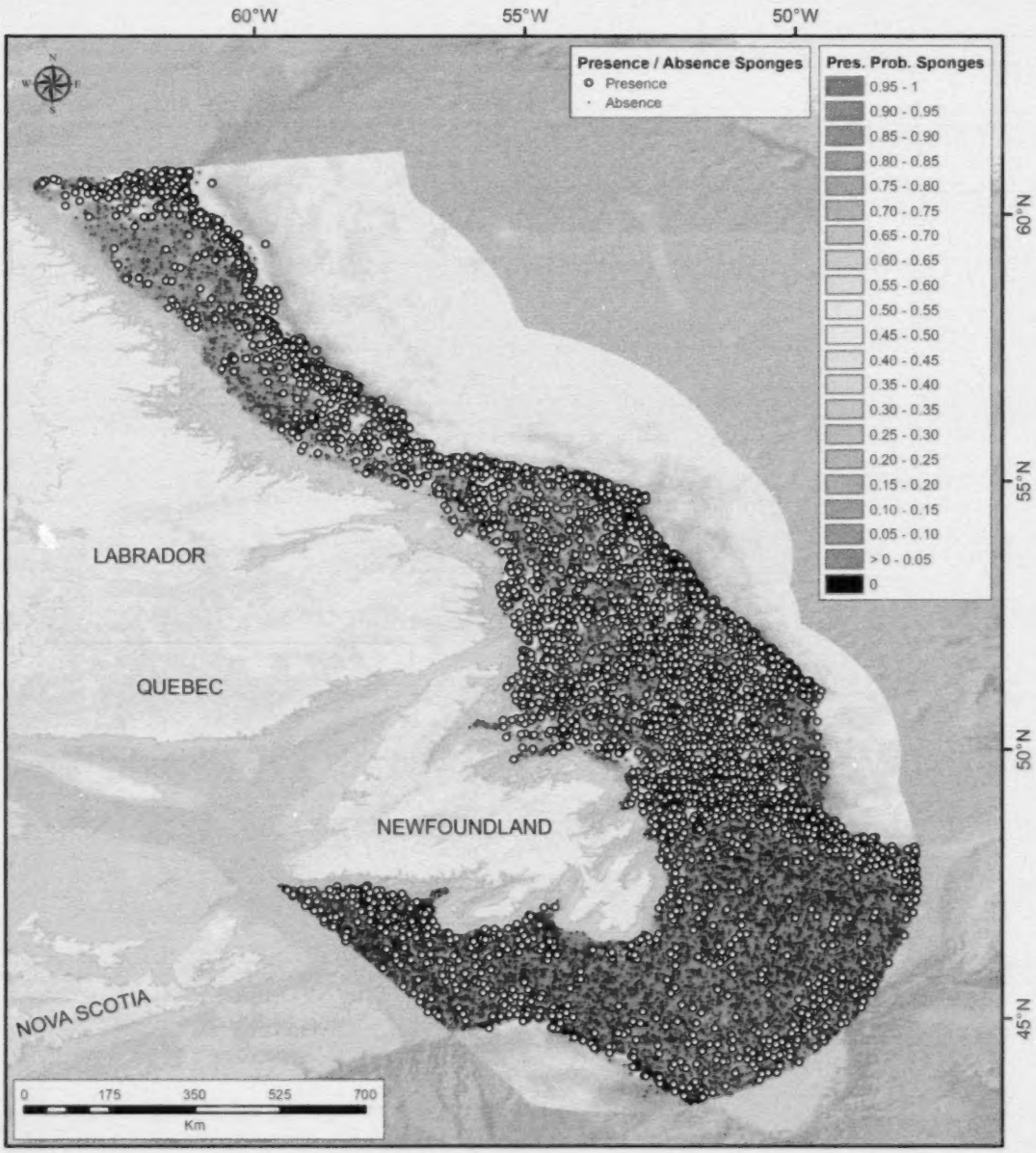

Figure 10. Presence and absence observations and predictions of presence probability of the unbalanced random forest model of sponge presence and absence data recorded from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 1995 and 2015.

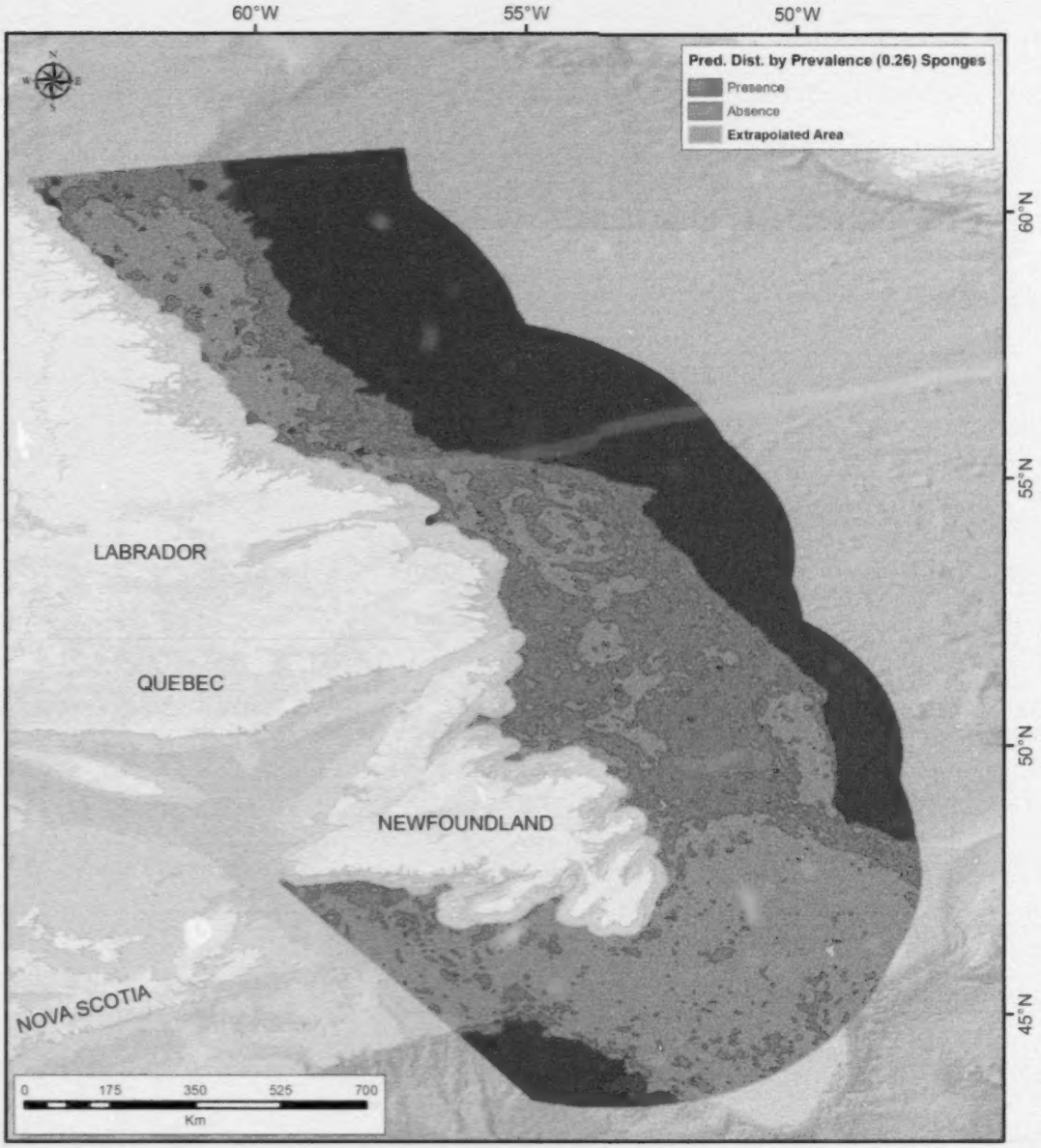

Figure 11. Predicted distribution (Pred. Dist.) of sponges in the Newfoundland and Labrador Region based on the prevalence threshold of 0.26 of sponge presence and absence data used in Model 2. Also shown are the areas of model extrapolation. Grey polygon may appear dark red or blue when overlain on the presence-absence surface.

Like Model 1, Fall Primary Production Average Maximum and Depth were the top two predictors of the sponge presence and absence data in Model 2 (Figure 12). Depth was followed by Fall Primary Production Average Minimum and Surface Temperature Average Maximum. Partial dependence plots for the top 6 predictor variables are shown in Figure 13. Along the Fall

Primary Production Average Maximum gradient the highest sponge presence probabilities occurred between 300 and 600 mg C m^{-2} day^{-1}. Along the Depth gradient, presence probability was highest between 1000 and 1500 m.

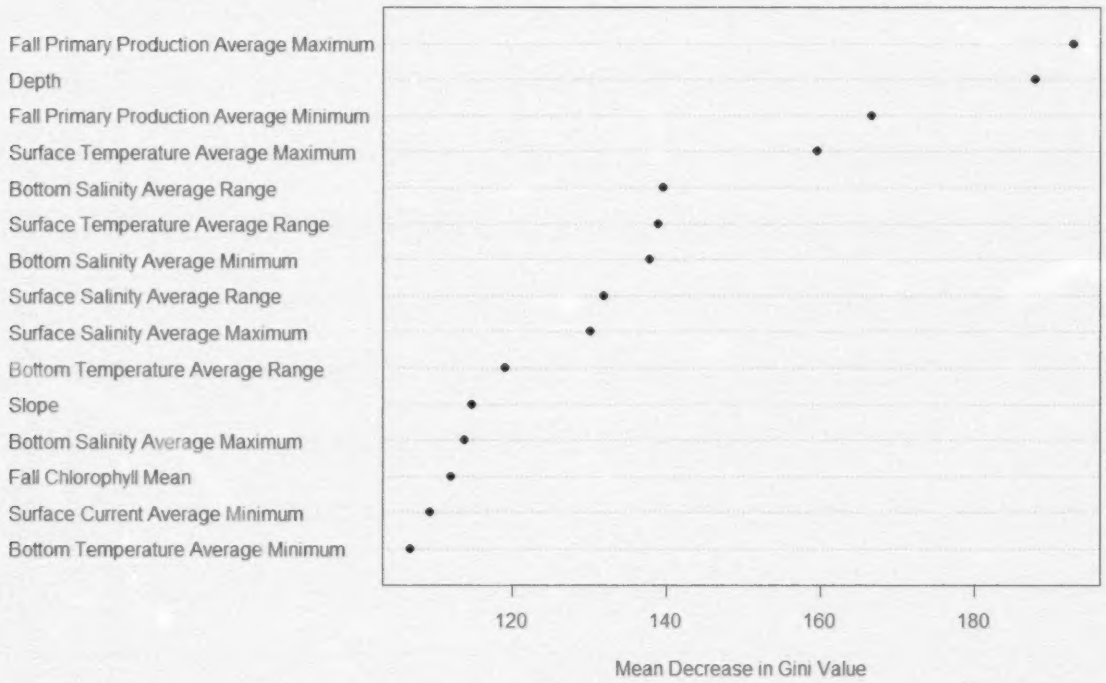

Figure 12. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the unbalanced random forest model of sponge presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini Value the more important the variable is for predicting the response data.

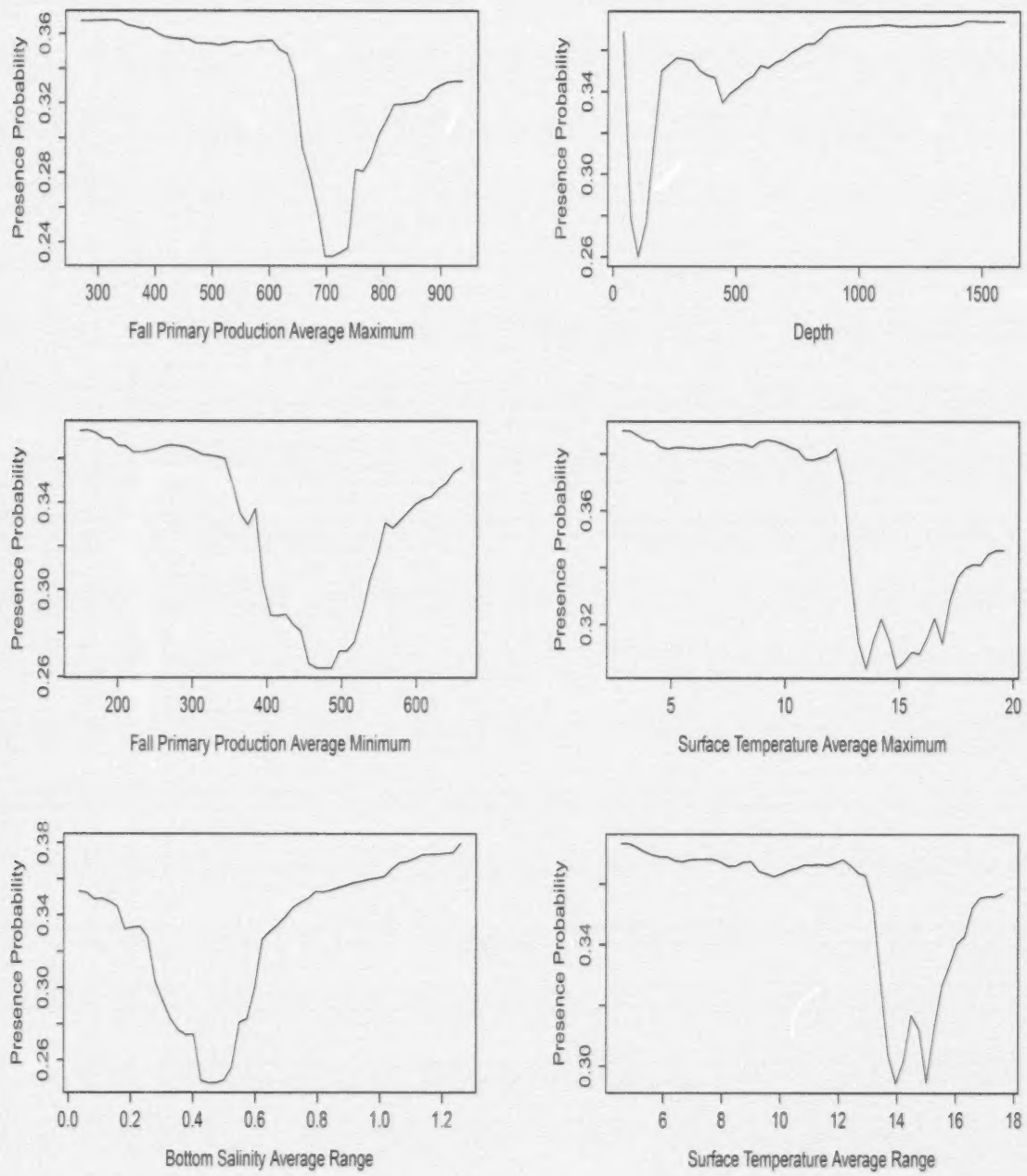

Figure 13. Partial dependence plots of the top 6 predictors from the random forest model of sponge presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The model using an unbalanced species prevalence and threshold equal to 0.26 (Model 2) was chosen as the best predictor of sponge distribution in the Newfoundland and Labrador Region. Although model accuracy measures were comparable between both models, Model 1 (balanced species prevalence) was considered a poor predictor of sponge presence probability due to its excessive extrapolation of high presence probabilities beyond the location of presence data.

Prediction of Sponge Biomass Using Random Forest

The accuracy measures of the regression random forest model on mean sponge biomass per grid cell are presented in Table 6. The highest R^2 value was 0.510, while the average was 0.360 ± 0.108 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.026 ± 0.006 SD. This model explained a high percentage of variance in the biomass data (average = 31.29% ± 1.92 SD).

Table 6. Accuracy measures from 10-fold cross validation of random forest model of average of sponge biomass (kg) per grid cell recorded from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 1995 and 2015. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.425	28.072	0.023	29.59
2	0.455	27.952	0.023	30.73
3	0.455	25.400	0.021	29.97
4	0.380	26.382	0.022	28.27
5	0.280	35.440	0.029	32.44
6	0.354	41.509	0.034	32.30
7	0.154	33.758	0.028	32.86
8	0.510	27.364	0.022	29.55
9	0.313	44.809	0.037	34.08
10	0.272	26.819	0.022	33.15
Mean	**0.360**	**31.756**	**0.026**	**31.29**
SD	**0.108**	**6.864**	**0.006**	**1.92**

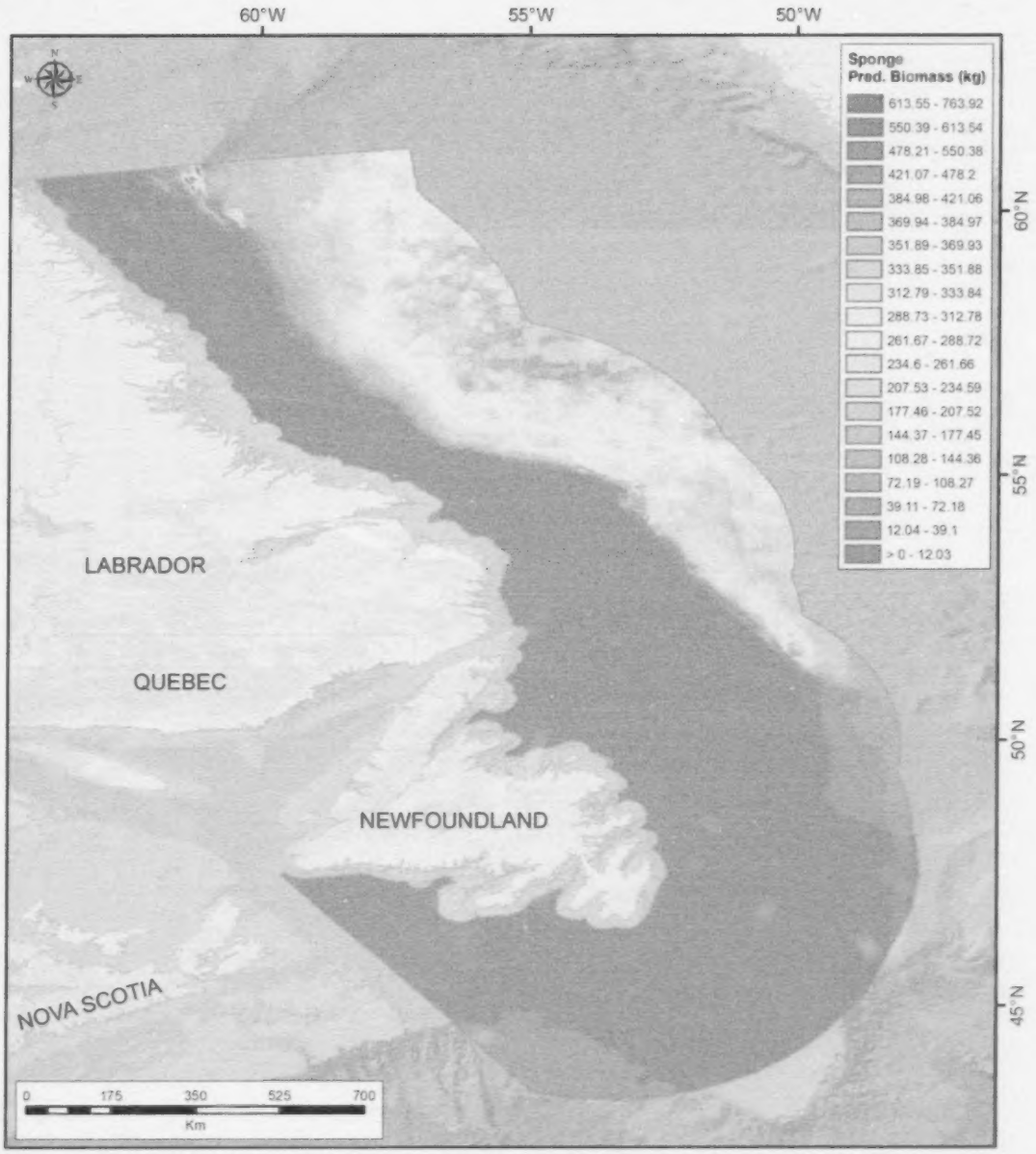

Figure 14. Predictions of biomass (kg) of sponges from catch recorded in DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 1995 and 2015.

Figures 14 and 15 show the predicted biomass surface of sponges. The entire shelf was predicted to have low (> 0 – 12.03 kg) sponge biomass. The slope off Saglek Bank had the highest predicted sponge biomass, reaching up to 763.92 kg. This area of high biomass was associated with a cluster of high mean catches (Figure 15). Interestingly, the slope off Hamilton Bank where a large cluster of high mean biomass catches occurred (Figure 15) appeared to be under-predicted by the model.

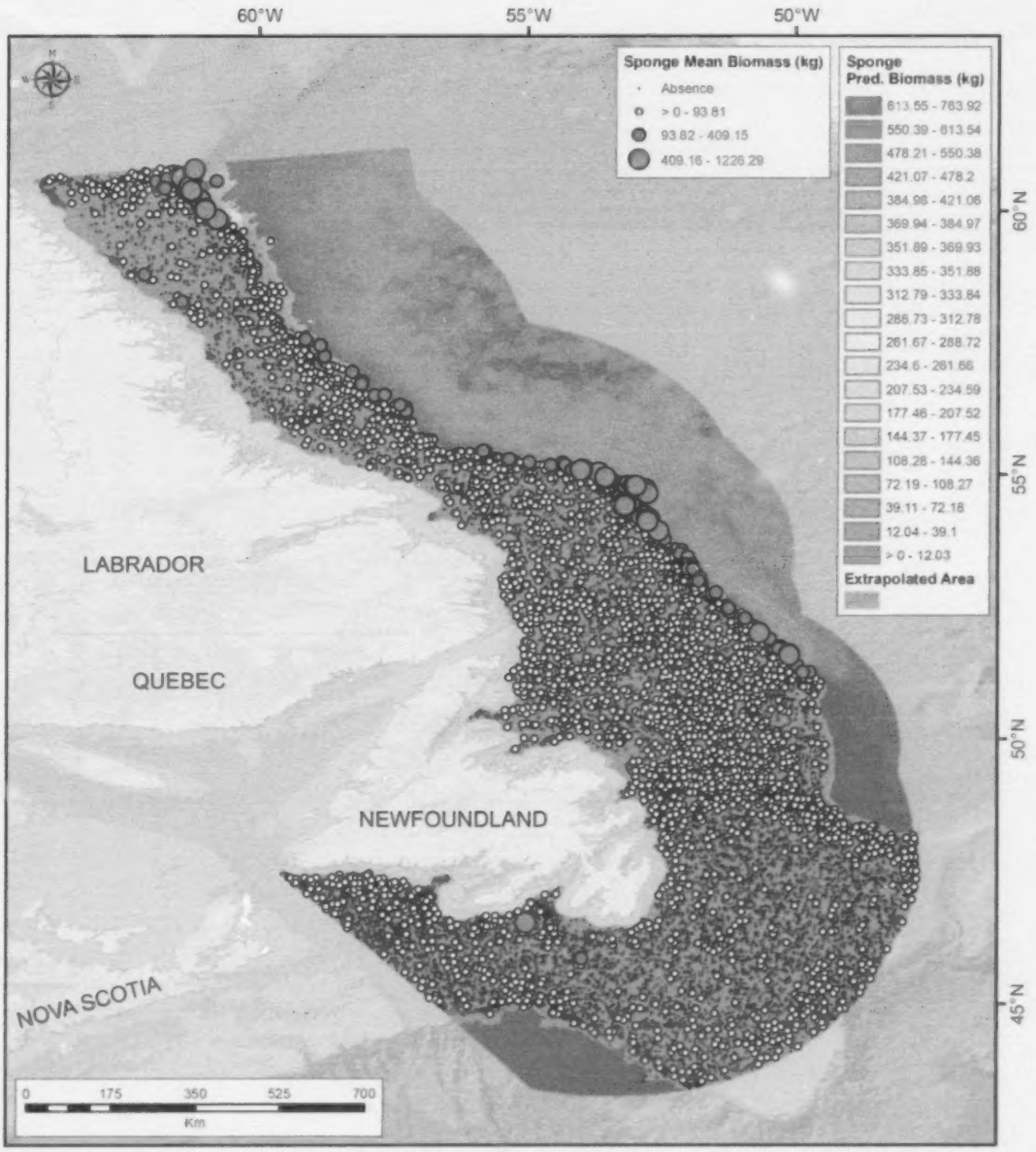

Figure 15. Predictions of biomass (kg) of sponges from catch recorded in DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 1995 and 2015. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

Of the 66 environmental variables used in the model, Summer Primary Production Average Minimum was the most important (Figure 16). This variable displayed a near-normal distribution prior to spatial interpolation (Guijarro et al., in prep.). Examination of the Q-Q plot revealed a spatial pattern to those data points over- and under-predicted by a normal distribution, with over-

predicted points located mainly in the northern portion of the study extent on Nain and Saglek Banks, and under-predicted points located on The Grand Banks of Newfoundland and the Northeast Newfoundland Shelf. Summer Primary Production Average Minimum was followed more distantly by Spring Primary Production Average Maximum, Bottom Salinity Average Range, and the remaining variables in the model. The partial dependence of sponge biomass on the top 6 most important variables is shown in Figure 17. Predicted biomass was highest at the lowest Summer Primary Production Average Minimum values (< 500 mg C m^{-2} day^{-1}). Values in this range coincided with those data points under-predicted by a normal distribution. The fit between predicted and observed values in the kriging model was good, with slight over-prediction of data points in that range. Some points could therefore be predicted higher than their true values and slightly outside the range of highest predicted biomass identified in the partial plot (Figure 17).

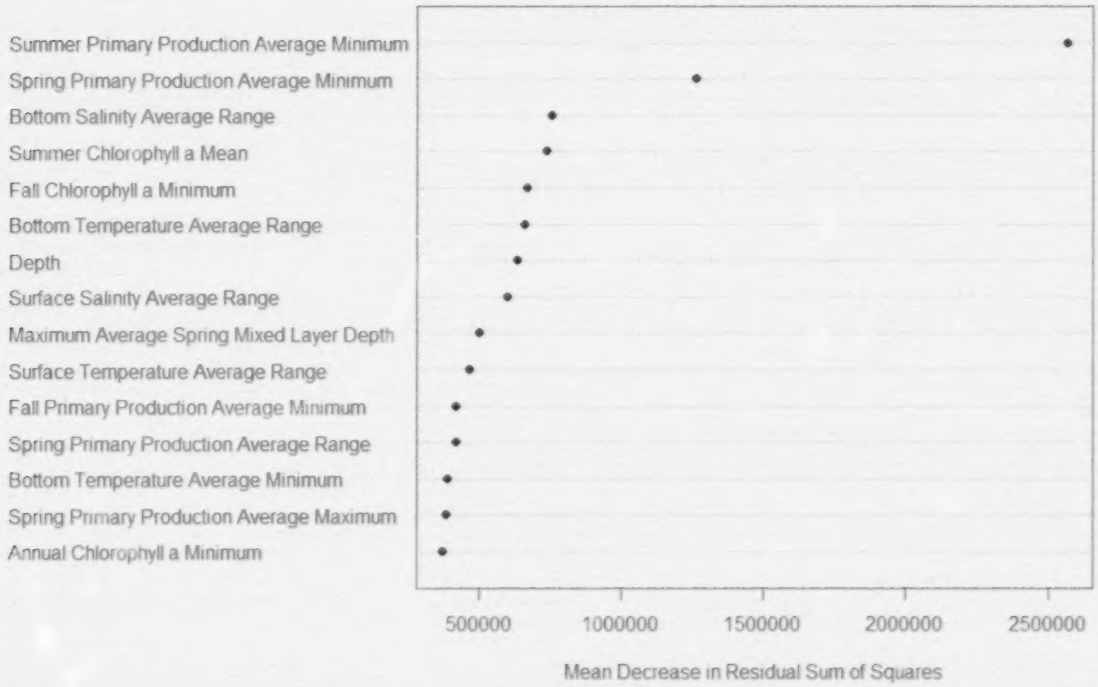

Figure 16. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sponge mean biomass data averaged per grid cell. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

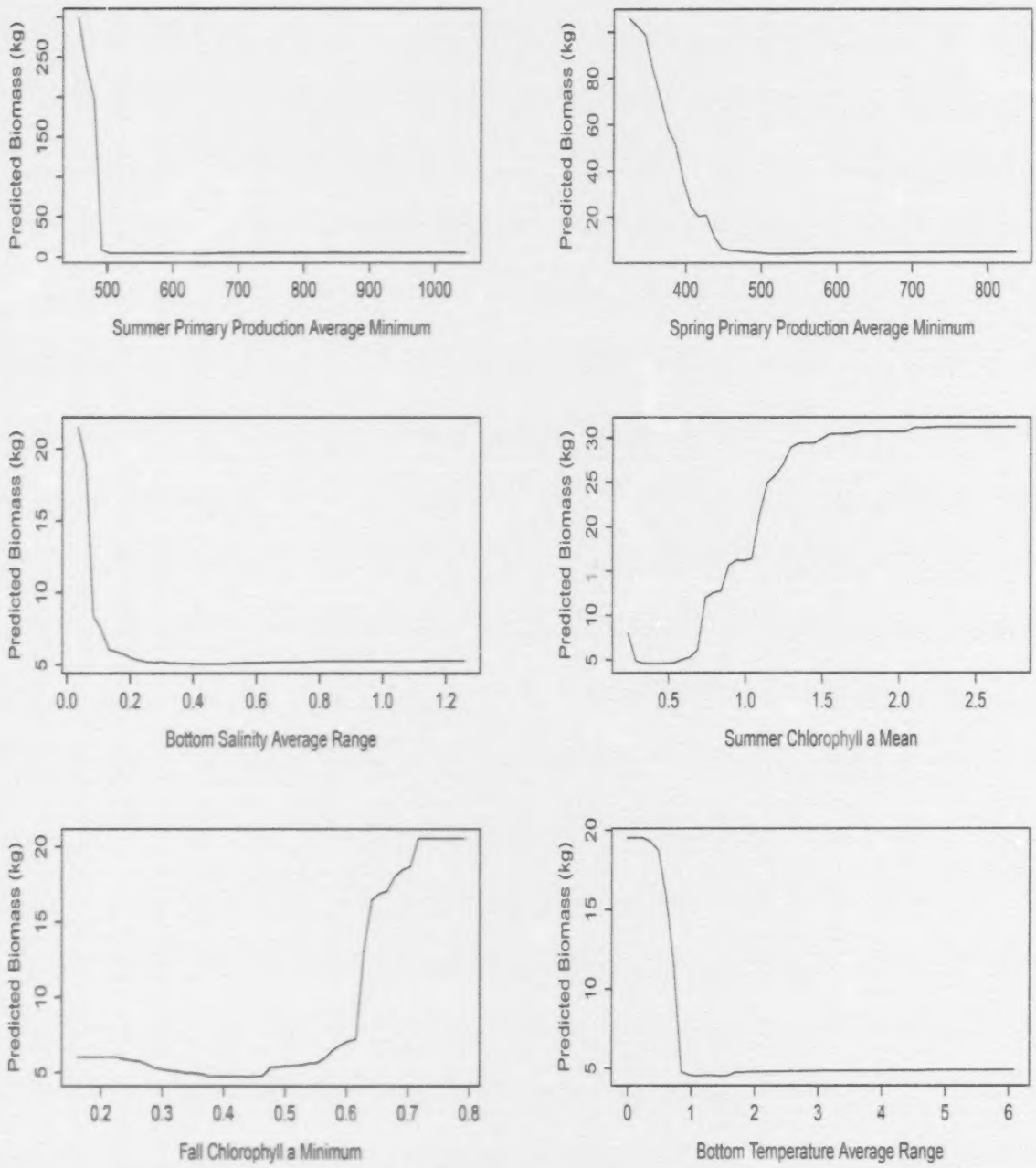

Figure 17. Partial dependence plots of the top six predictors from the random forest model of sponge biomass data collected within the Newfoundland and Labrador Region between 1995 and 2015, ordered left to right from the top. Predicted biomass is shown on the *y*-axis of each graph.

Sea Pens (Pennatulacea)

Data Sources and Distribution

Sea pen catch data was collected over a span of 13 years from 2003 to 2015 and consisted of 946 presence and 4773 absence records (Table 7). Absence records were distributed relatively evenly across the study extent (Figure 18). However, presence records had a highly uneven distribution and were concentrated on the slopes off Newfoundland and off Nain and Saglek Banks in northern Labrador. The highest mean biomass records (up to 40 kg) were located in the Laurentian Channel. A single large catch also occurred on Nain Bank off Labrador.

Table 7. Number of presence and absence records of sea pens catch recorded from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

Year	Total of number of presences	Total of number of absences
2003	1	81
2004	9	225
2005	17	275
2006	54	442
2007	113	430
2008	100	406
2009	115	411
2010	101	593
2011	94	478
2012	95	451
2013	101	478
2014	85	361
2015	61	142

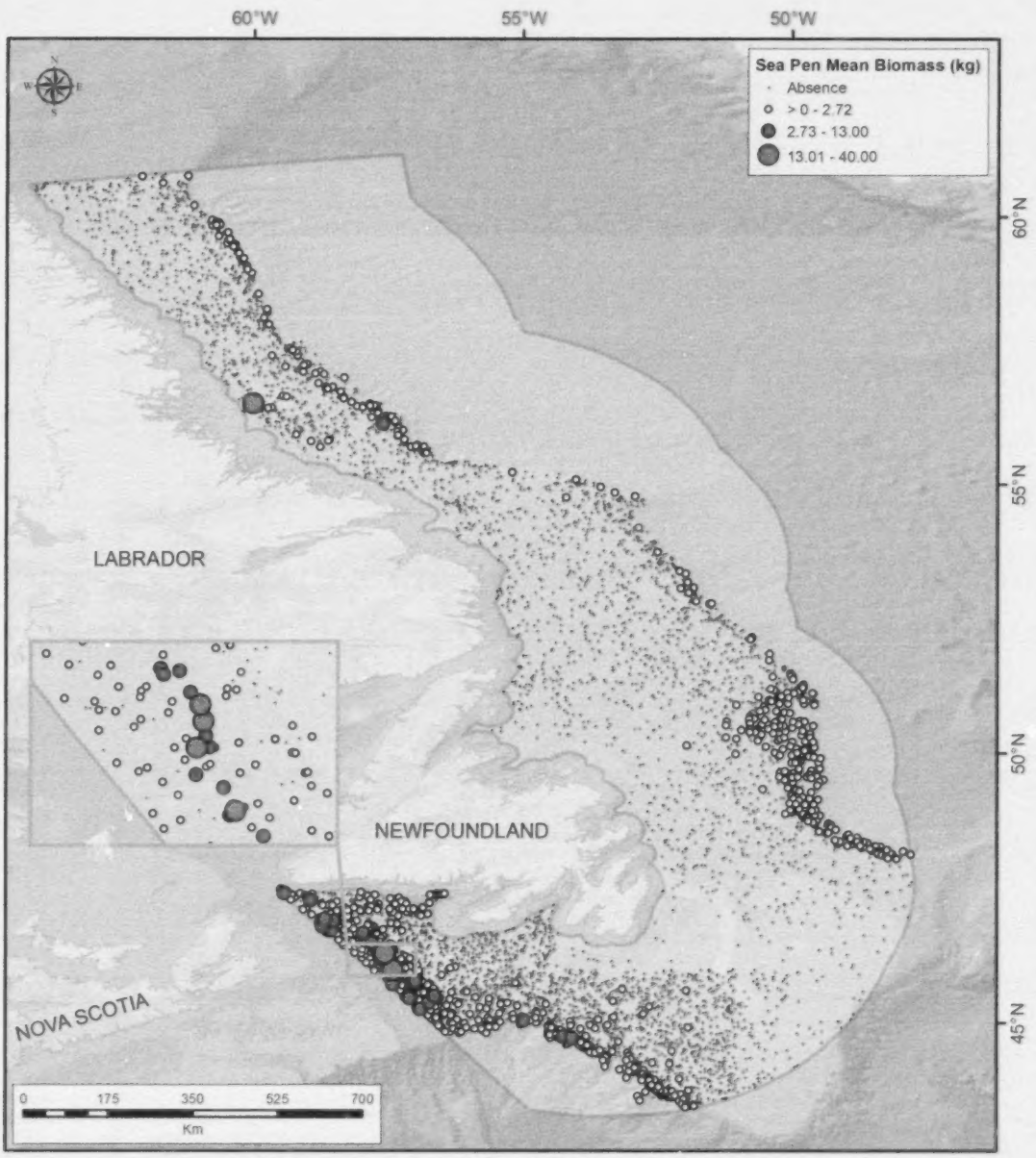

Figure 18. Mean biomass (kg) per grid cell of sea pen catch data recorded from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

Model 1 – Balanced Species Prevalence

Accuracy measures (mean AUC, sensitivity, and specificity) for the random forest model on balanced species prevalence (946 presences and 946 absences; Model 1) are presented in Table 8. The average AUC was 0.928, indicating excellent model performance. The highest mean AUC of 0.935 was associated with Model run 10 and is therefore considered the optimal model for the prediction of the sea pen response data. The sensitivity and specificity measures of this model were 0.865 and 0.850, respectively. The confusion matrix of the optimal model is also presented in Table 8. Class error for both the presence and absence classes was low (0.135 and 0.150, respectively).

Table 8. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of sea pens within the Newfoundland and Labrador Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 10) which is considered the optimal model for predicting the presence probability of sea pens.

Model Run	AUC	Sensitivity	Specificity
1	0.932	0.867	0.848
2	0.927	0.864	0.852
3	0.927	0.868	0.853
4	0.930	0.854	0.856
5	0.927	0.856	0.844
6	0.921	0.845	0.832
7	0.927	0.851	0.840
8	0.925	0.862	0.836
9	0.930	0.856	0.851
10	**0.935**	**0.865**	**0.850**
Mean	**0.928**	**0.859**	**0.846**
SD	**0.004**	**0.008**	**0.008**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	804	142	946	0.150
Presence	128	818	946	0.135

The presence probability prediction surface of sea pens is presented in Figure 19. The highest predictions of presence probability occurred in the Laurentian Channel and along the southwest and northeast slopes of the Grand Banks of Newfoundland. The slope off Nain was also predicted to have moderate to high presence probability of sea pens. In general the shelf was predicted to have low presence probability of sea pens. Areas of high presence probability corresponded well with the spatial distribution of presence records (Figure 20), with little extrapolation beyond the location of these data points.

The actual presence and absence data observations (946 presences and 946 absences) used in the optimal run of Model 1 showed some slight spatial bias across the study area (Figure 21). Also shown in this figure are the areas of model extrapolation. All deep water beyond the continental slope was considered extrapolated, with a few small pockets distributed across the shelf.

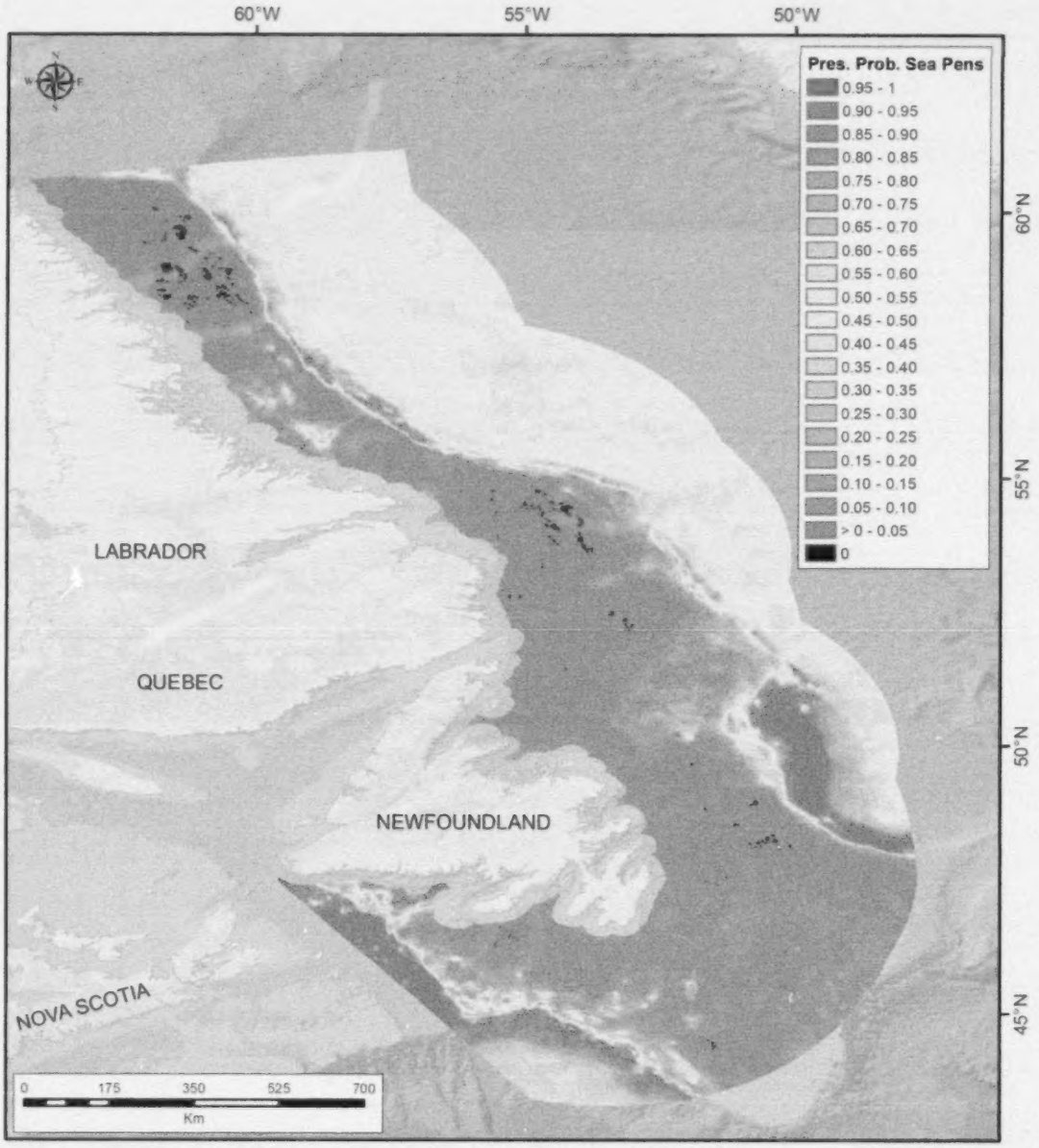

Figure 19. Predictions of presence probability from the optimal random forest model of sea pen presence and absence data collected from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

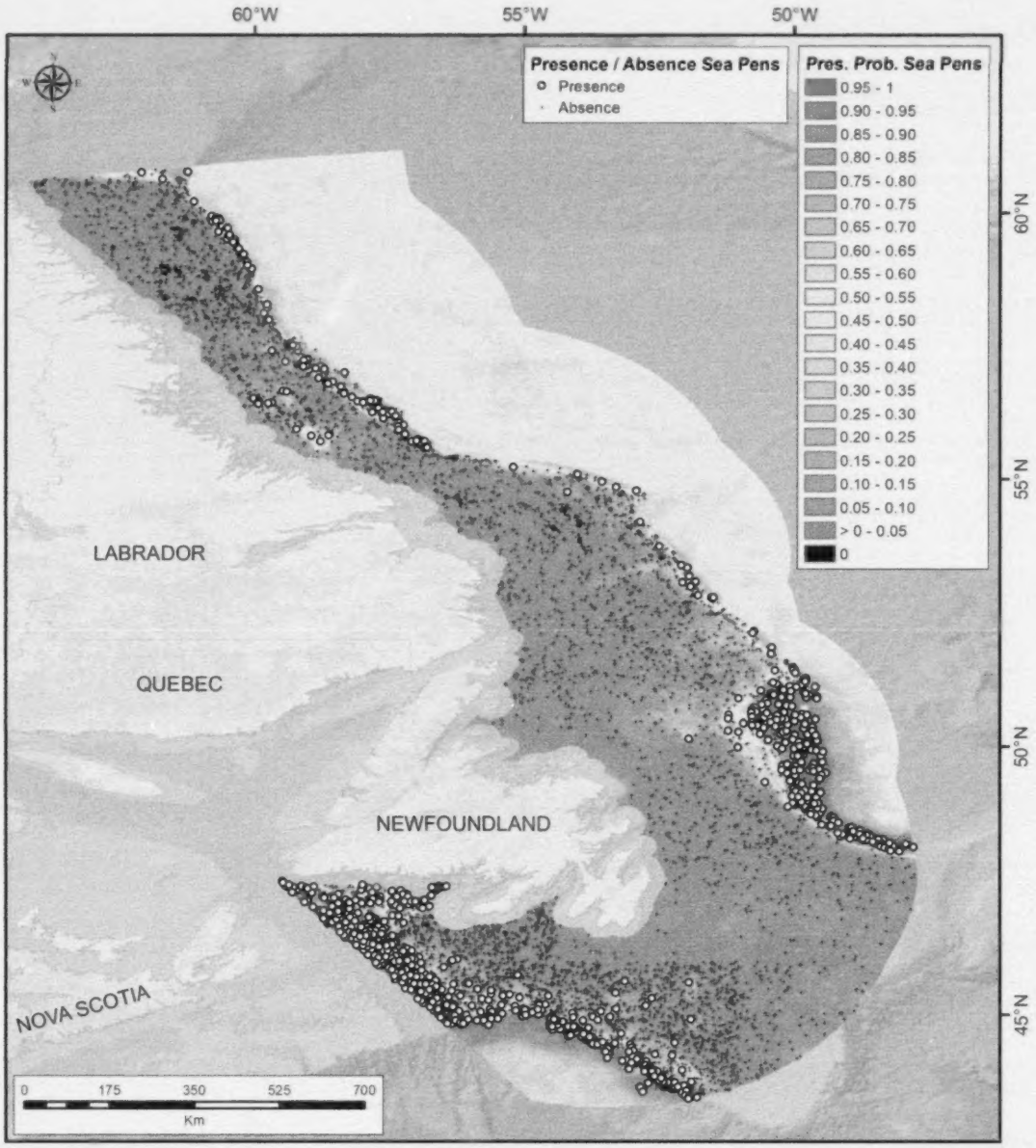

Figure 20. Presence and absence observations and predictions of presence probability of the optimal random forest model of sea pen presence and absence data collected from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

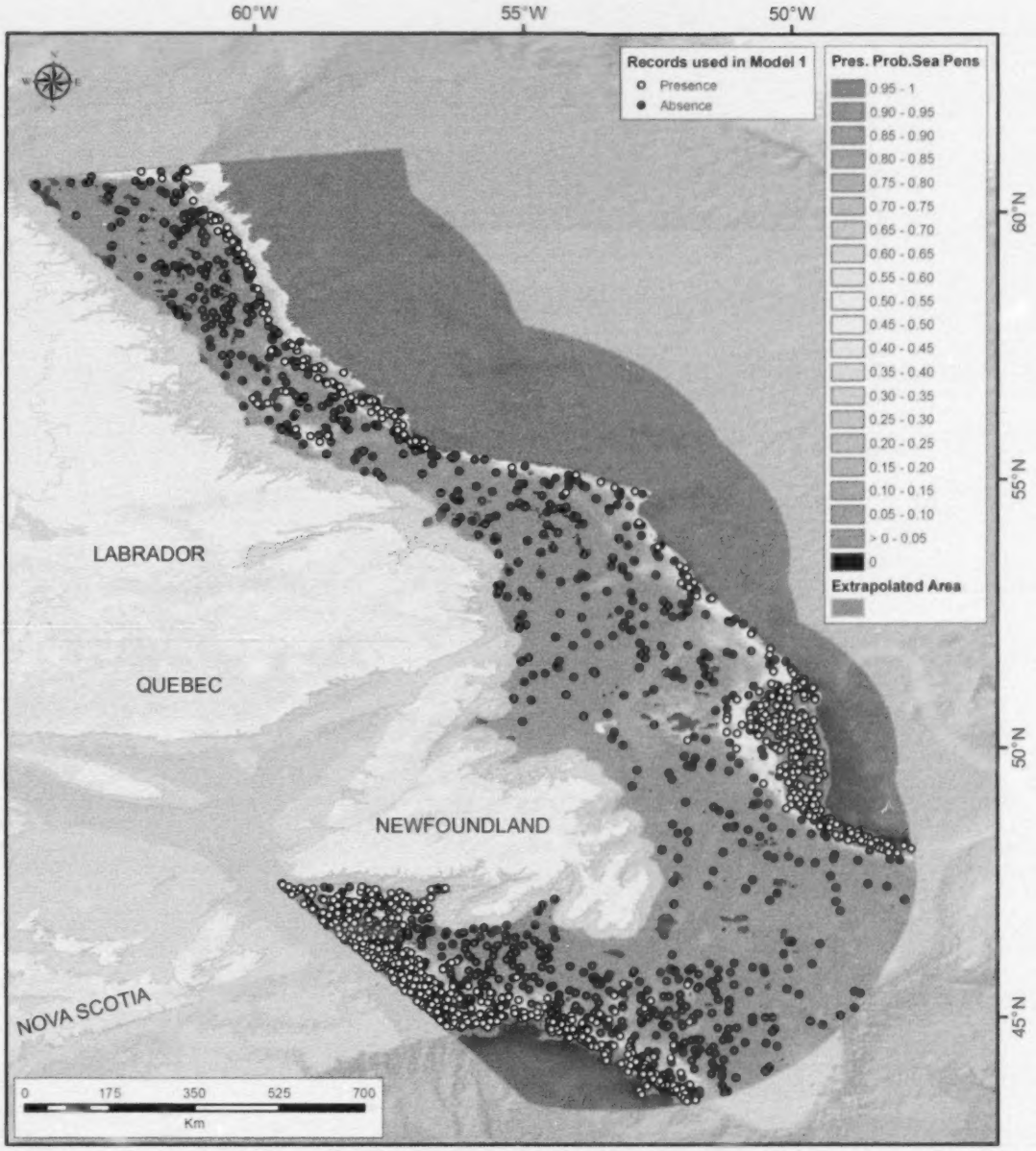

Figure 21. Map of the 1892 data observations (946 presences and 946 absences) of sea pens used in the optimal random forest Model 1. Also shown is the predicted presence probability (pres. prob.) of sea pens generated from Model 1 and areas of model extrapolation.

Of all 66 environmental predictor variables used in the model, Bottom Temperature Average Minimum was the most important for the classification of the sea pen presence and absence data (Figure 22). Prior to spatial interpolation, this variable displayed a bimodal distribution (Guijarro et al., in prep.). Examination of the Q-Q plot revealed a spatial pattern to the data points over- and under-predicted by a normal distribution, with over-predicted points located mainly on The

Northeast Newfoundland Shelf and in the deep waters off the Labrador Slope, and under-predicted points located along the coast, on Grand Bank, and along a narrow band over the slopes of Newfoundland and Labrador. Bottom Temperature Average Minimum was followed by Bottom Salinity Average Maximum, Bottom Salinity Average Minimum, and Depth. Partial dependence plots for the top 6 predictor variables are shown in Figure 23. Presence probability of sea pens was highest between 3 and 6°C along the gradient in Bottom Temperature Average Minimum. Values in this range coincided with under-predicted points located along a narrow band over the slopes of Newfoundland and Labrador. The fit between predicted and observed values in the kriging model was fair, with under-prediction of values between 3 and 6°C. Some points could therefore be predicted lower than their true values and slightly outside the range of the highest predicted presence probability identified in the partial plot (Figure 23).

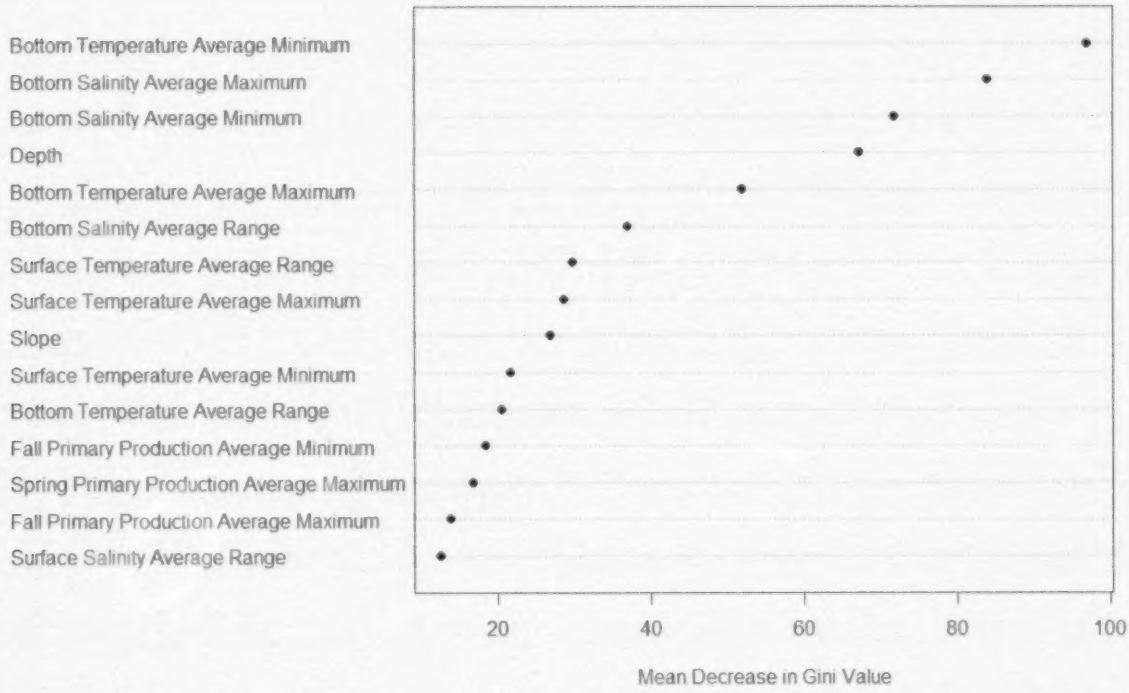

Figure 22. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting sea pen presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini value the more important the variable is for predicting the response data.

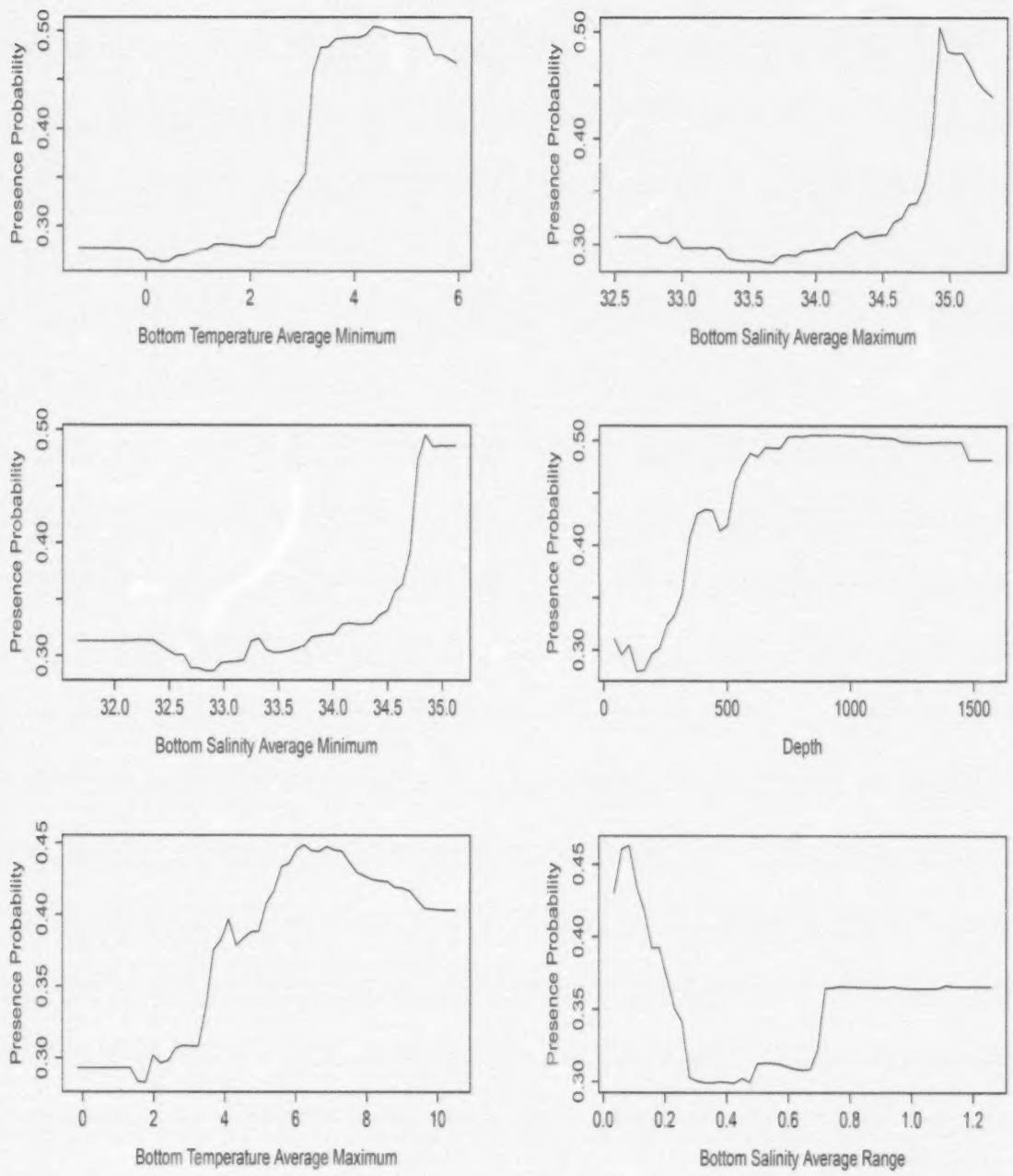

Figure 23. Partial dependence plots of the top 6 predictors from the optimal random forest model of sea pen presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 9 shows the accuracy measures for the random forest model using all sea pen presence and absence data (946 presences and 4773 absences; Model 2) and a threshold equal to species prevalence (0.17). The average AUC calculated from this model was slightly lower than that of Model 1 (0.926 compared to 0.928 of Model 1). Sensitivity and specificity were also slightly lower than Model 1.

The surface of predicted presence probability of sea pens generated from Model 2 is presented in Figure 24. The areas of high predicted presence probability identified in Model 1 were much reduced in Model 2, particularly along the slope off Nain Bank. There was good spatial congruence between areas of high presence probability and presence observations, with little extrapolation beyond these data points (Figure 25). Figure 26 depicts the classification of sea pen presence probability into presence and absence categories based on the prevalence threshold of 0.17. In this map, all presence probability values generated from Model 2 that were greater than 0.17 were classified as presence, while values less than 0.17 were classed as absence. Much of the continental shelf was classified as absence of sea pens. The Laurentian Channel and much of the slopes were classified as presence of sea pens.

Table 9. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model presence and absence of sea pens within the Newfoundland and Labrador Region between 2003 and 2015. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.937		**Absence**	**Presence**				
2	0.921	**Absence**	4030	743	4773	0.156	0.847	0.844
3	0.927	**Presence**	145	801	946	0.153		
4	0.925							
5	0.925							
6	0.914							
7	0.932							
8	0.940							
9	0.927							
10	0.913							
Mean	**0.926**							
SD	**0.009**							

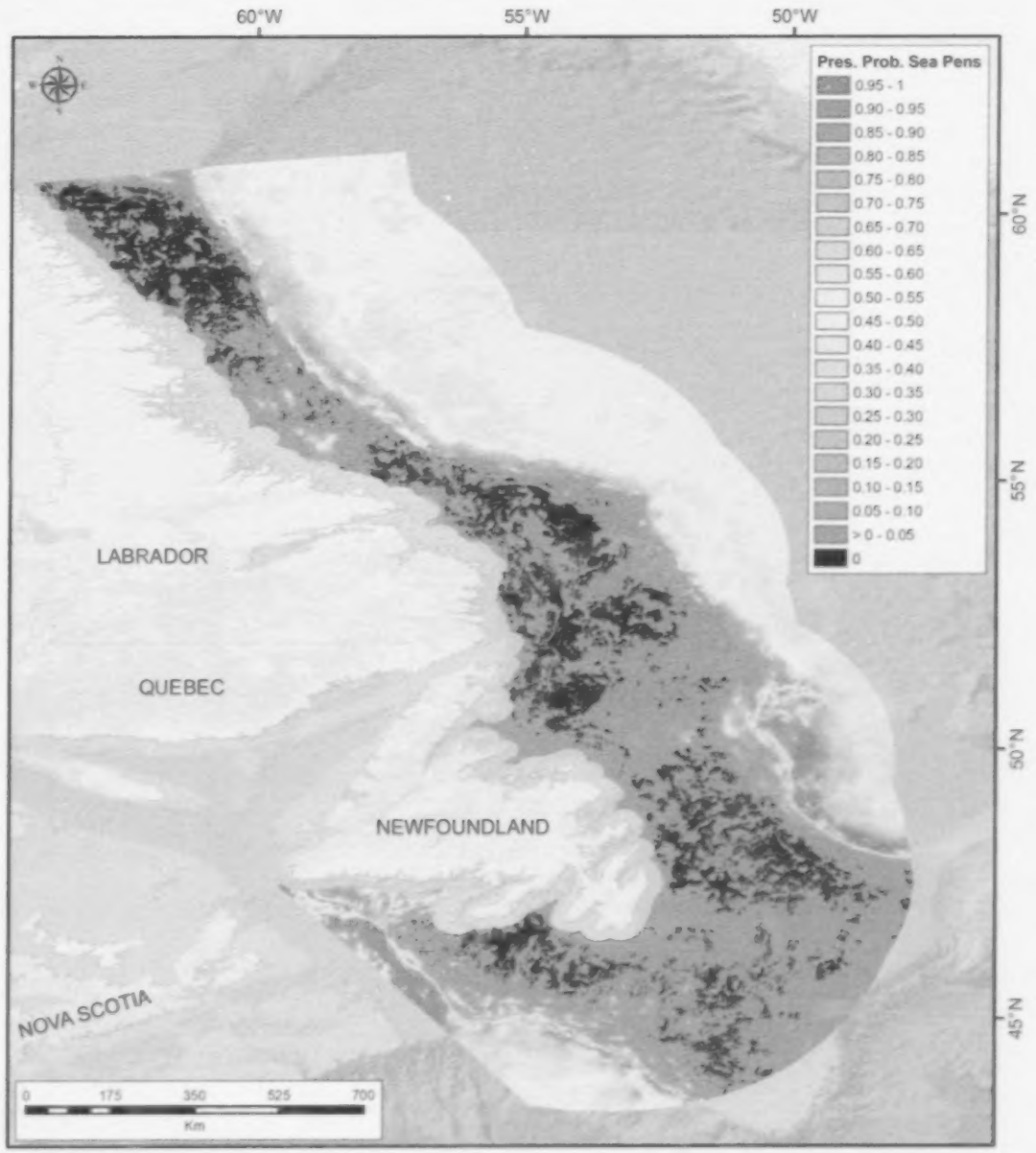

Figure 24. Predictions of presence probability from the unbalanced random forest model of sea pen presence and absence data collected from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

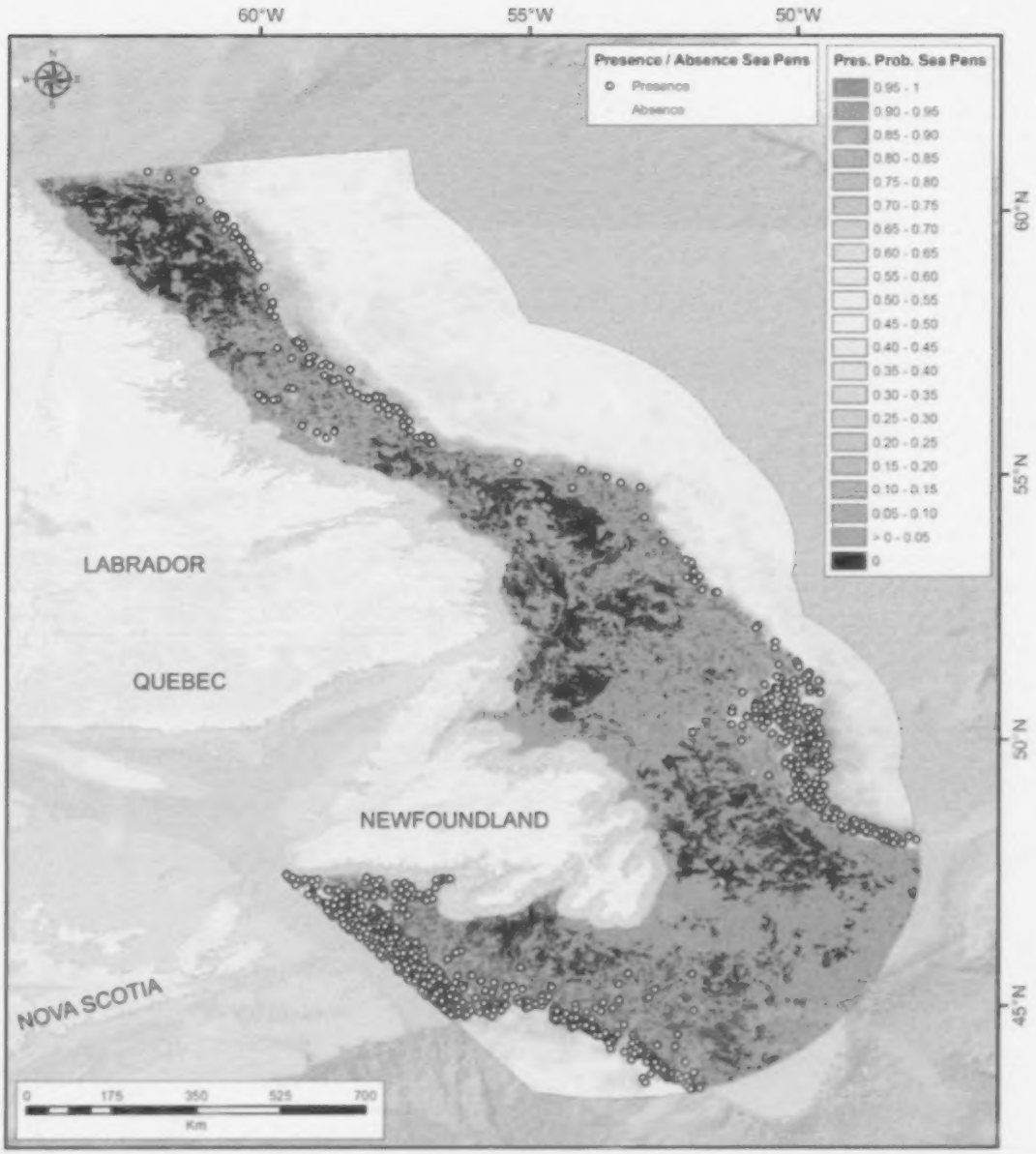

Figure 25. Presence and absence observations and predictions of presence probability from the unbalanced random forest model of sea pens presence and absence data collected from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

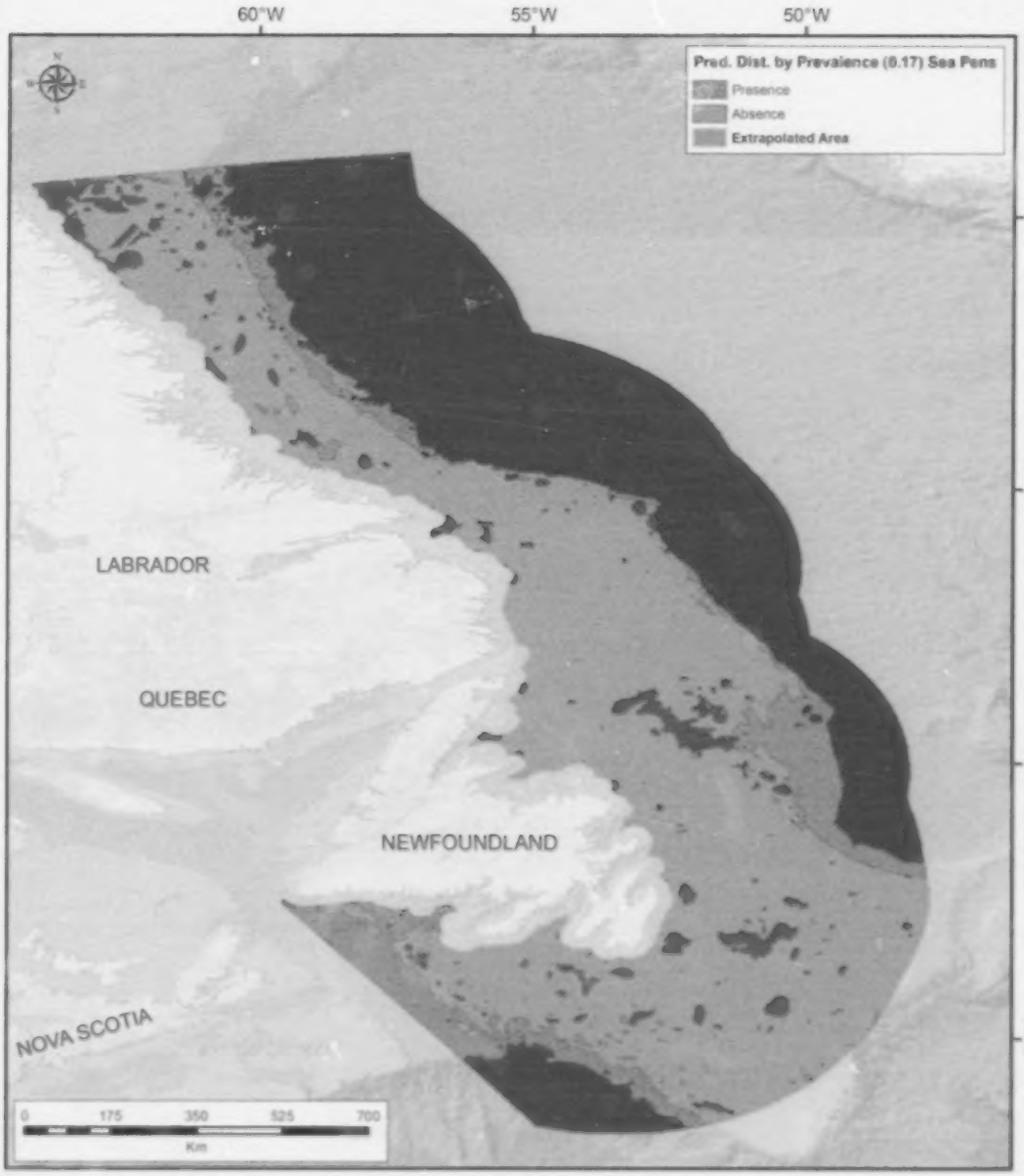

Figure 26. Predicted distribution (Pred. Dist.) of sea pens in the Newfoundland and Labrador Region based on the prevalence threshold of 0.17 of sea pen presence and absence data used in Model 2. Also shown are the areas of model extrapolation. Grey polygon may appear dark red or blue when overlain on the presence-absence surface.

The importance of the environmental predictor variables for predicting the presence probability of sea pens catch data is presented in Figure 27. Depth (a non-interpolated variable) was the most important variable for the classification of the sea pen presence and absence data. This variable was followed by Bottom Temperature Average Minimum, Bottom Salinity Average Minimum,

and Bottom Salinity Average Maximum. Partial dependence of the sea pen presence and absence data on the top 6 predictor variables is shown in Figure 28. Sea pen presence probability was highest between 500 m to 1500 m along the depth gradient. Along the gradient in Bottom Temperature Average Minimum, presence probability rapidly increased at ~3°C and remained high.

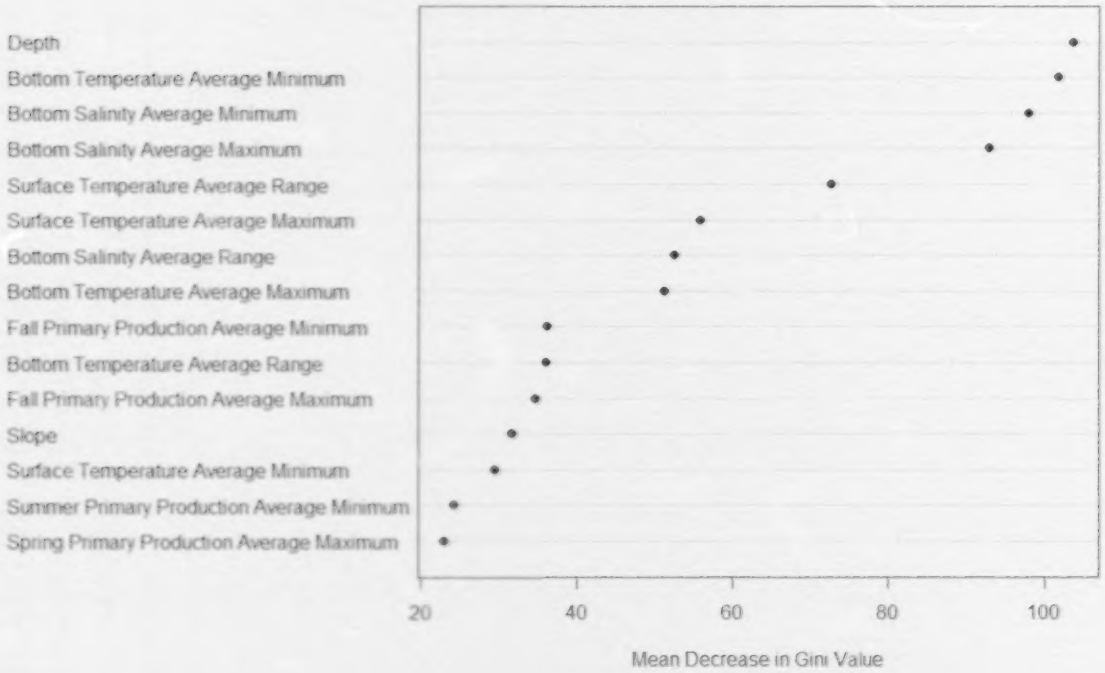

Figure 27. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the unbalanced random forest model of sea pen presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini Value the more important the variable is for predicting the response data.

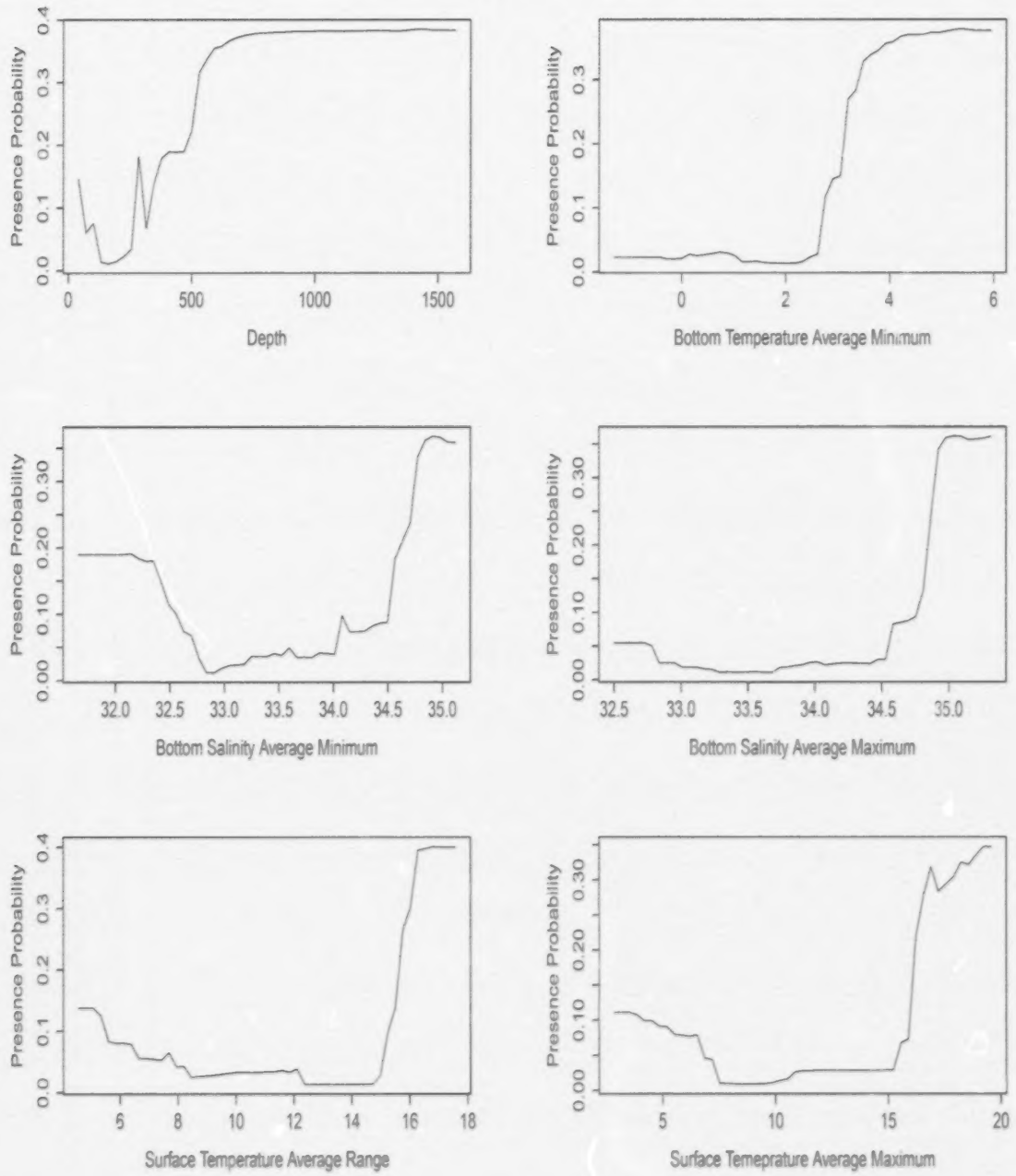

Figure 28. Partial dependence plots of the top 6 predictors from the unbalanced random forest model of sea pen presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model Selection

The random forest model using all available sea pen records and an unbalanced species prevalence and threshold equal to 0.17 (Model 2) was chosen as the best predictor of sea pen distribution in the Newfoundland and Labrador Region. Model 1 (balanced species prevalence) was considered a poor predictor of presence probability of sea pens due to its exaggeration of high presence probability beyond the location of presence data, particularly in the Laurentian Channel and slope off the Northeast Newfoundland Shelf. This phenomenon was likely due to random down-sampling of absence data.

Prediction of Sea Pen Biomass Using Random Forest

The accuracy measures of the regression random forest model on mean sea pen biomass per grid cell are presented in Table 10. The highest R^2 value was 0.642, while the average was 0.376 ± 0.202 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.018 ± 0.010 SD. This model explained a moderate percentage of variance in the biomass data (average = 28.74% ± 3.25 SD).

Table 10. Accuracy measures from 10-fold cross validation of random forest model of average of sea pen biomass (kg) per grid cell recorded from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.642	0.494	0.012	24.73
2	0.475	1.479	0.037	25.58
3	0.210	0.466	0.012	28.45
4	0.099	0.466	0.012	31.28
5	0.391	0.440	0.011	29.70
6	0.405	0.490	0.012	31.15
7	0.438	1.049	0.026	26.81
8	0.048	1.251	0.031	35.40
9	0.410	0.447	0.011	28.27
10	0.640	0.503	0.013	25.99
Mean	**0.376**	**0.708**	**0.018**	**28.74**
SD	**0.202**	**0.394**	**0.010**	**3.25**

Figures 29 and 30 show the predicted biomass surface of sea pens. The majority of the study extent was predicted to have low (> 0 – 0.1 kg) sea pen biomass. The highest predicted biomass

(up to 24.27 kg) occurred in a small area in the Laurentian Channel. This area of high biomass was associated with the cluster of high biomass values in that area (Figure 30).

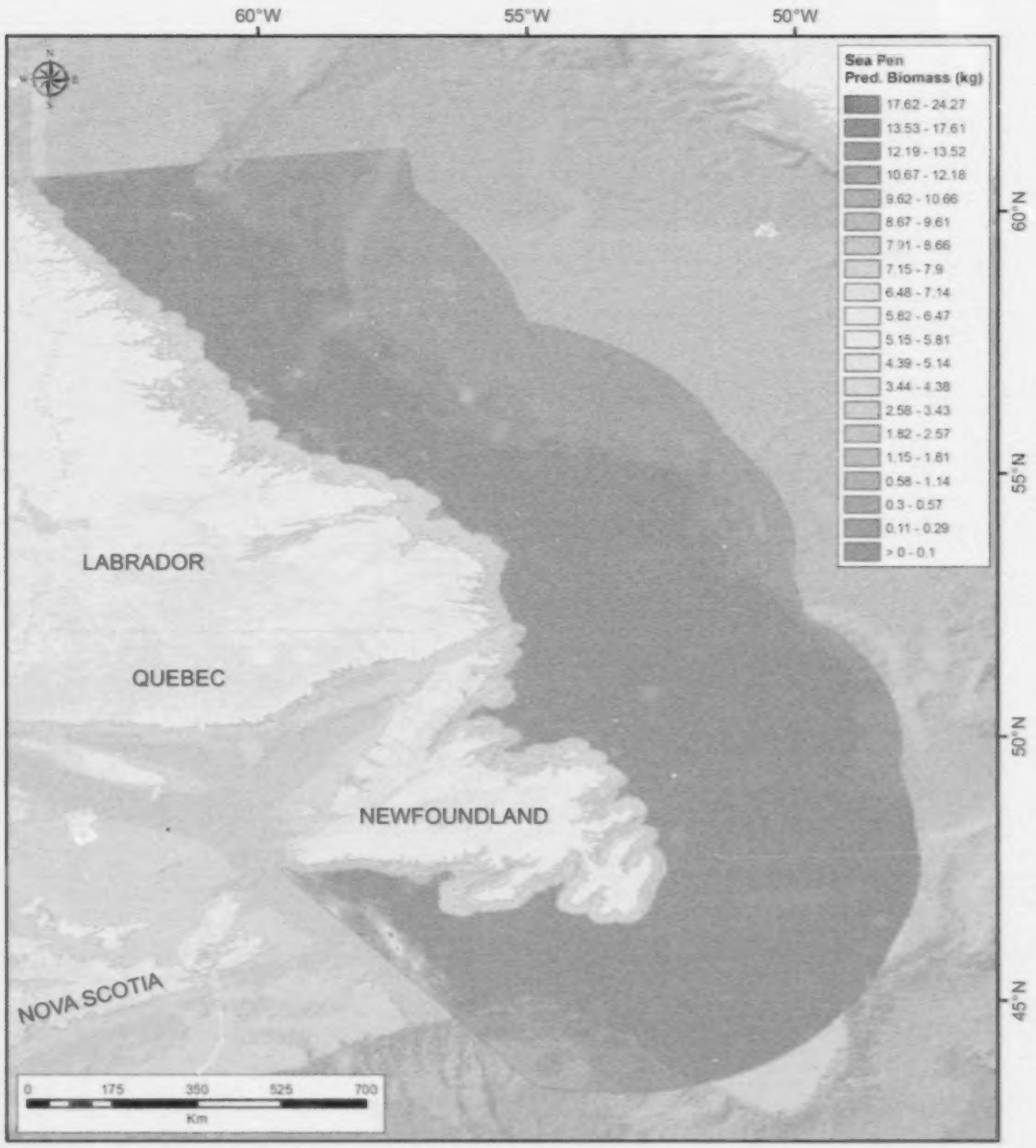

Figure 29. Predictions of biomass (kg) of sea pens from catch recorded in DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

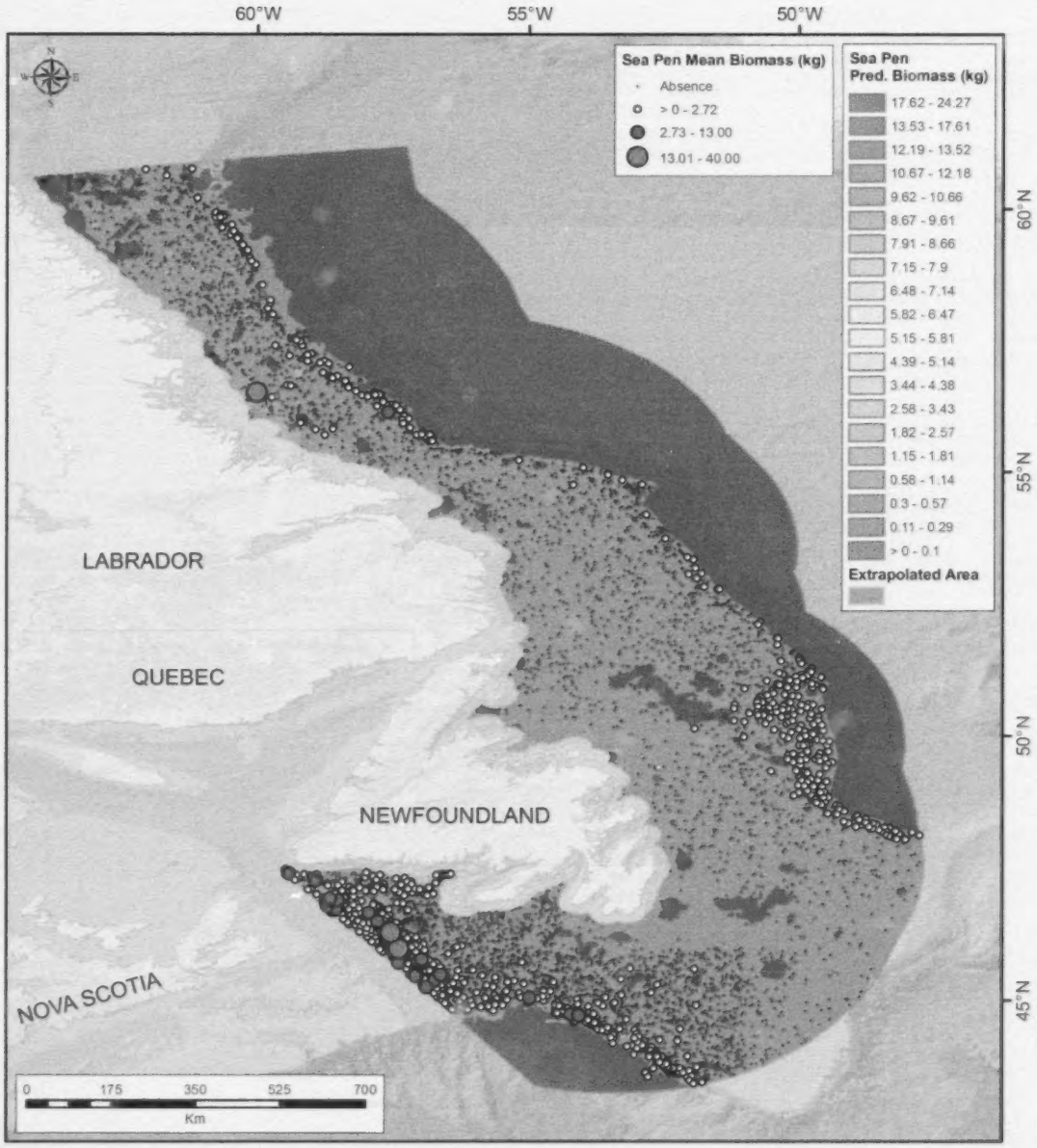

Figure 30. Predictions of biomass (kg) of sea pens from catch recorded in DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting sea pen biomass are shown in Figure 31. Maximum Average Winter Mixed Layer Depth was the most important variable in the model. Prior to spatial interpolation, this variable displayed a highly right-skewed distribution (Guijarro et al., in prep.). Examination of the Q-Q plot revealed a strong spatial pattern to those

data points over- and under-predicted by a normal distribution, with over-predicted points located along the coast and in deep waters off Nain and Saglek Banks, and under-predicted points located on Saglek Bank, in deep waters off Hamilton Bank and the Northeast Newfoundland Shelf, and along the slopes of Labrador. Maximum Average Winter Mixed Layer Depth was followed more distantly by Fall Primary Production Average Range and Surface Temperature Average Range. The partial dependence of sea pen biomass on the top 6 most important variables is shown in Figure 32. Predicted biomass was highest at the lowest Maximum Average Winter Mixed Layer Depth values (< 30 m), and then sharply decreased between ~ 30 and 50 m and then rapidly increased and plateaued at ~50 m. Values less than 30 m coincided with those over-predicted points along the Labrador coast, while data greater than 50 m corresponded to both over- and under-predicted points in the deep water off the Labrador shelf. These values are not of particular concern however, as the fit between predicted and observed values in the kriging model was excellent, with only slight over-prediction at low values.

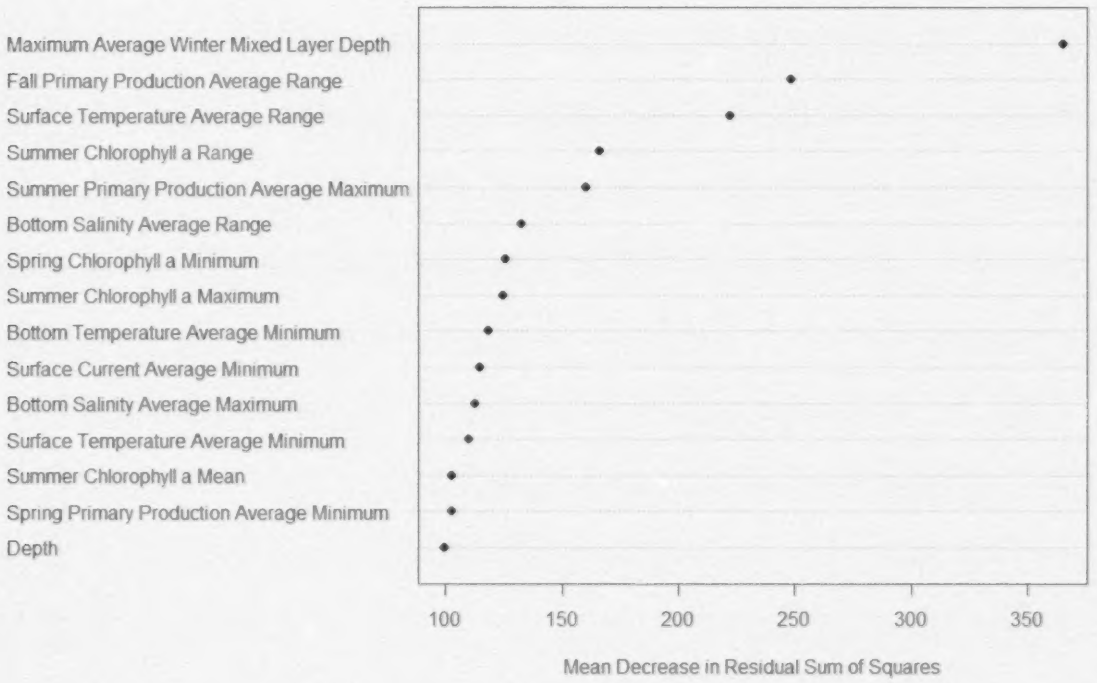

Figure 31. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sea pen mean biomass data averaged per grid cell. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

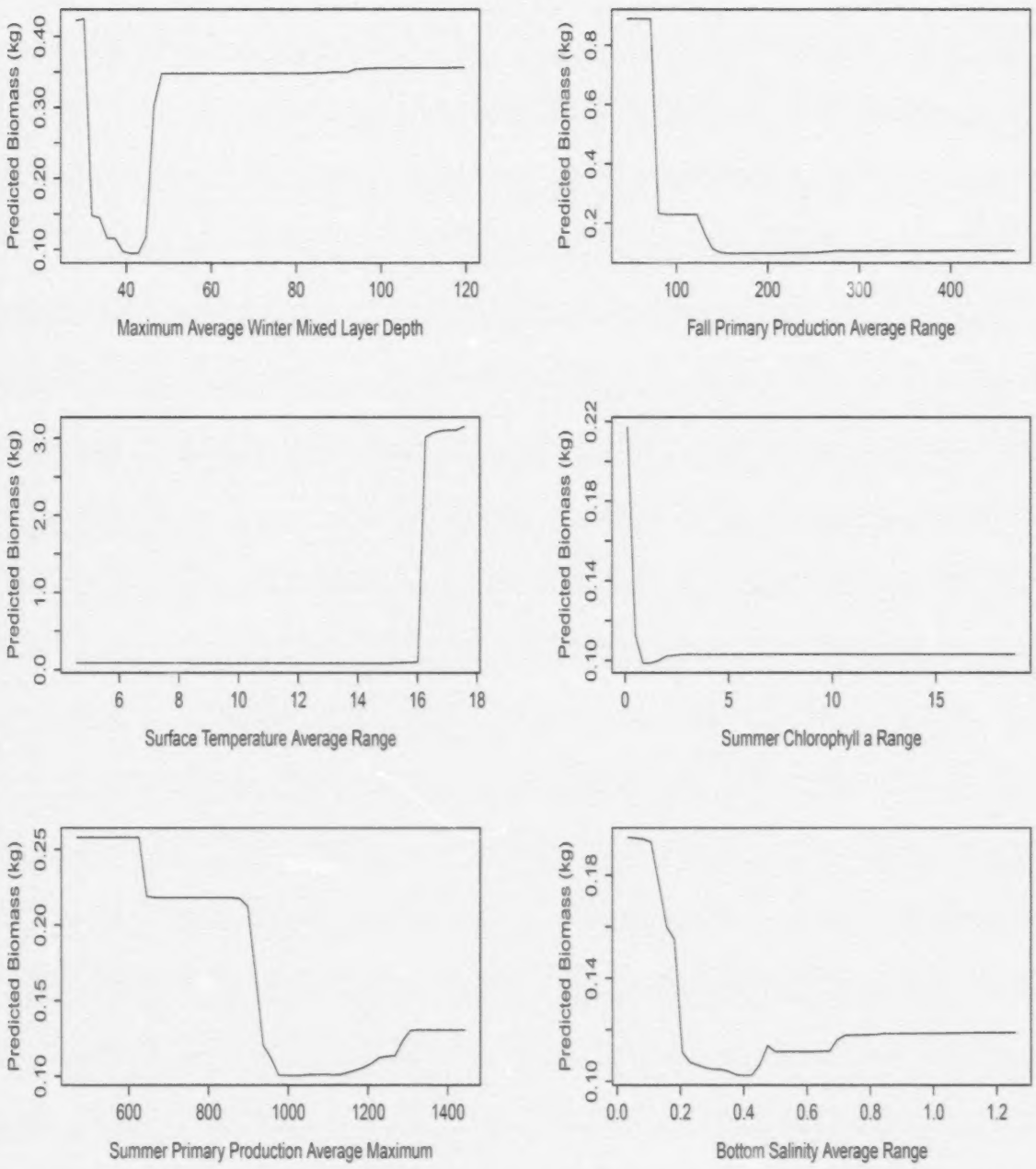
Predicted Biomass (kg)
0.10 0.20 0.30 0.40
40 60 80 100 120
Maximum Average Winter Mixed Layer Depth
Predicted Biomass (kg)
0.2 0.4 0.6 0.8
100 200 300 400
Fall Primary Production Average Range
Predicted Biomass (kg)
0.0 1.0 2.0 3.0
6 8 10 12 14 16 18
Surface Temperature Average Range
Predicted Biomass (kg)
0.10 0.14 0.18 0.22
0 5 10 15
Summer Chlorophyll a Range
Predicted Biomass (kg)
0.10 0.15 0.20 0.25
600 800 1000 1200 1400
Summer Primary Production Average Maximum
Predicted Biomass (kg)
0.10 0.14 0.18
0.0 0.2 0.4 0.6 0.8 1.0 1.2
Bottom Salinity Average Range

Figure 32. Partial dependence plots of the top six predictors from the random forest model of sea pen biomass data collected within the Newfoundland and Labrador Region between 2003 and 2015, ordered left to right from the top. Predicted biomass is shown on the *y*-axis of each graph.

Large Gorgonian Corals

Data Sources and Distribution

Large gorgonian coral catch data was collected over a span of 13 years from 2003 to 2015 and consisted of 514 presence and 5651 absence records (Table 11). Absence records were distributed relatively evenly across the study extent (Figure 33). However, presence records had a highly uneven distribution and were concentrated mainly along the slopes of Newfoundland and Labrador, although some large gorgonian records were scattered across the shelf. The highest mean biomass records (up to 288.97 kg) were located on the slope off Saglek Bank off northern Labrador.

Table 11. Number of presence and absence records of large gorgonian coral catch recorded from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

Year	Total of number of presences	Total of number of absences
2003	8	74
2004	17	151
2005	24	284
2006	50	556
2007	47	501
2008	53	380
2009	58	594
2010	59	674
2011	49	563
2012	65	586
2013	40	522
2014	40	642
2015	4	124

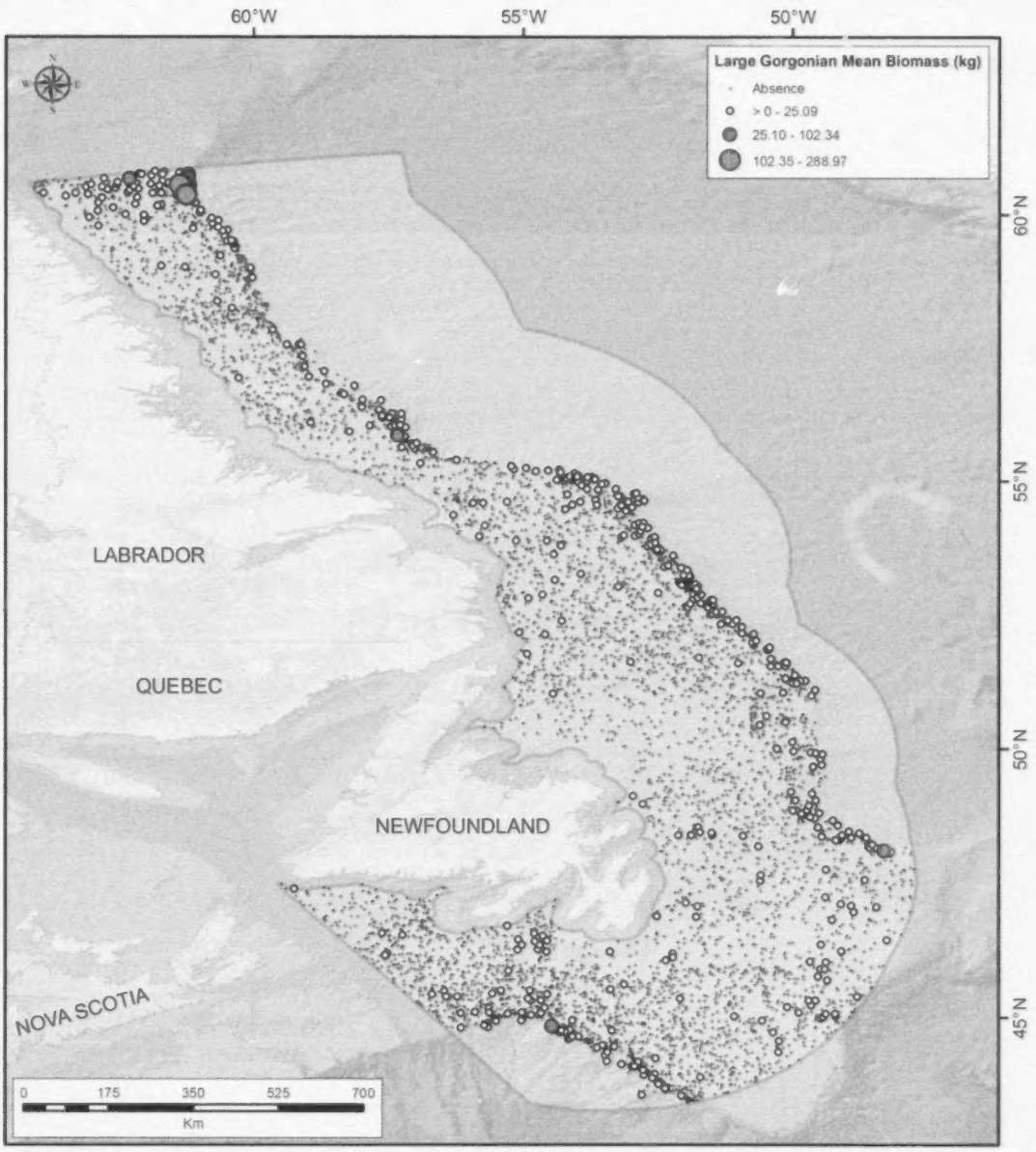

Figure 33. Mean biomass (kg) per grid cell of large gorgonian coral catch data recorded from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

Model 1 – Balanced Species Prevalence

Accuracy measures (mean AUC, sensitivity, and specificity) for the random forest model on balanced species prevalence (514 presences and 514 absences; Model 1) are presented in Table 12. The average AUC was 0.8110, indicating good model performance. The highest mean AUC of 0.826 was associated with Model run 9 and is therefore considered the optimal model for the prediction of the large gorgonian coral response data. The sensitivity and specificity measures of this model were 0.722 and 0.790, respectively. The confusion matrix of the optimal model is also presented in Table 12. Class error for both the presence and absence classes was somewhat moderate (0.385 and 0.210, respectively).

Table 12. Accuracy measures for all 10 model repetitions of 10 fold cross validation of a random forest model of presence and absence of large gorgonian corals within the Newfoundland and Labrador Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 9) which is considered the optimal model for predicting the presence probability of large gorgonian corals.

Model Run	AUC	Sensitivity	Specificity
1	0.798	0.677	0.761
2	0.825	0.720	0.767
3	0.777	0.691	0.732
4	0.826	0.728	0.784
5	0.819	0.706	0.790
6	0.801	0.687	0.749
7	0.791	0.693	0.737
8	0.826	0.737	0.776
9	**0.826**	**0.722**	**0.790**
10	0.821	0.712	0.772
Mean	**0.811**	**0.707**	**0.769**
SD	**0.018**	**0.020**	**0.021**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	406	108	514	0.210
Presence	143	371	514	0.385

The presence probability prediction surface of large gorgonian corals from Model 1 is presented in Figure 34. The highest predictions of presence probability occurred along of the northeast slope of Newfoundland and the Labrador Slope. Northern Saglek Bank had a high predicted presence probability of large gorgonian corals. These areas of high presence probability corresponded well with the spatial distribution of presence records (see Figure 35) although some extrapolation of high presence probabilities occurred beyond the location of presence data.

The actual presence and absence data observations (514 presences and 514 absences) used in the optimal run of Model 1 showed some slight spatial bias across the study area, particularly along the slope (Figure 36). Also shown in this figure are the areas of model extrapolation. Deep water beyond the slope was considered extrapolated area. Smaller pockets of extrapolated area are distributed across the shelf and in coastal areas.

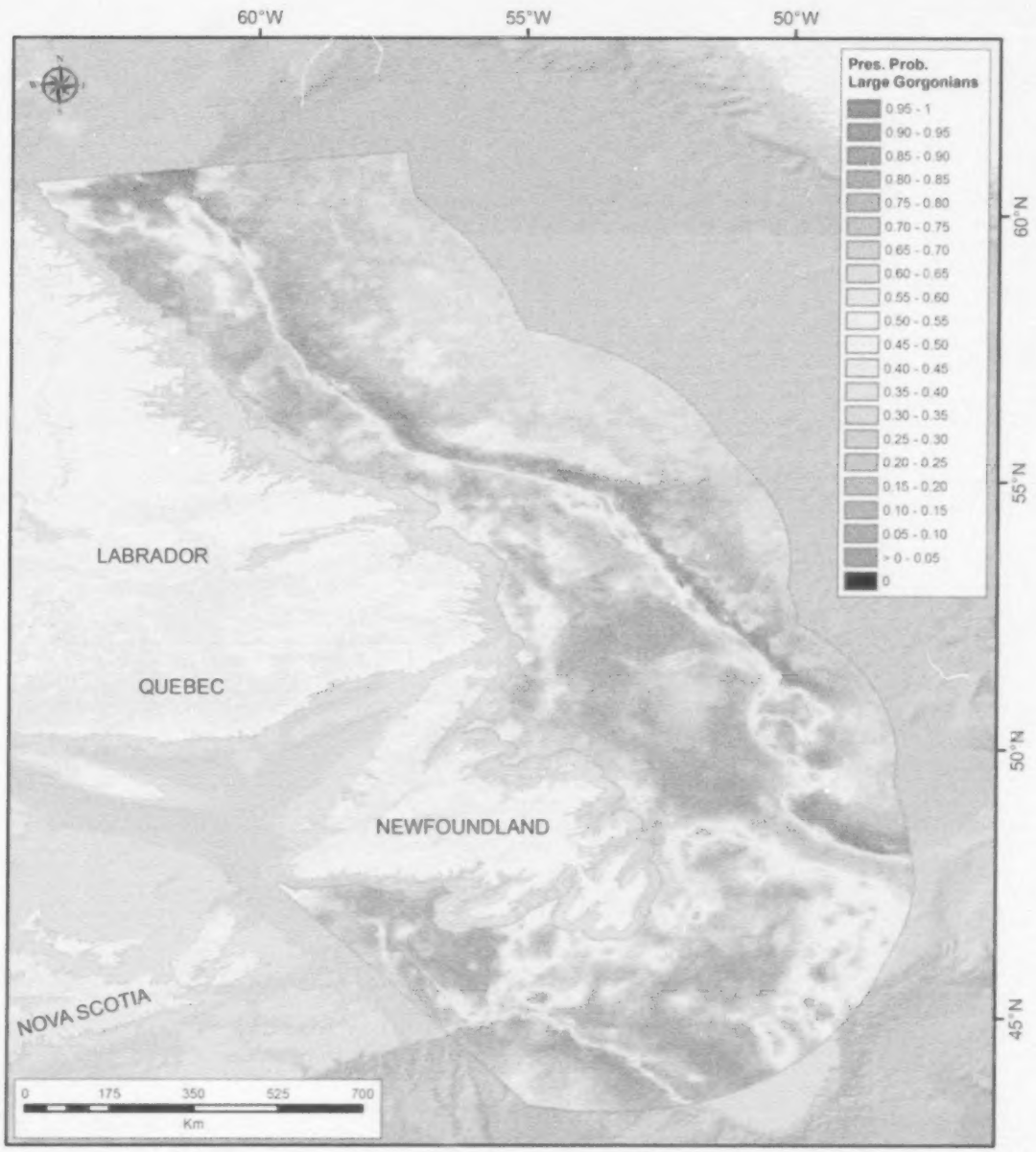

Figure 34. Predictions of presence probability from the optimal random forest model of large gorgonian coral presence and absence data collected from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

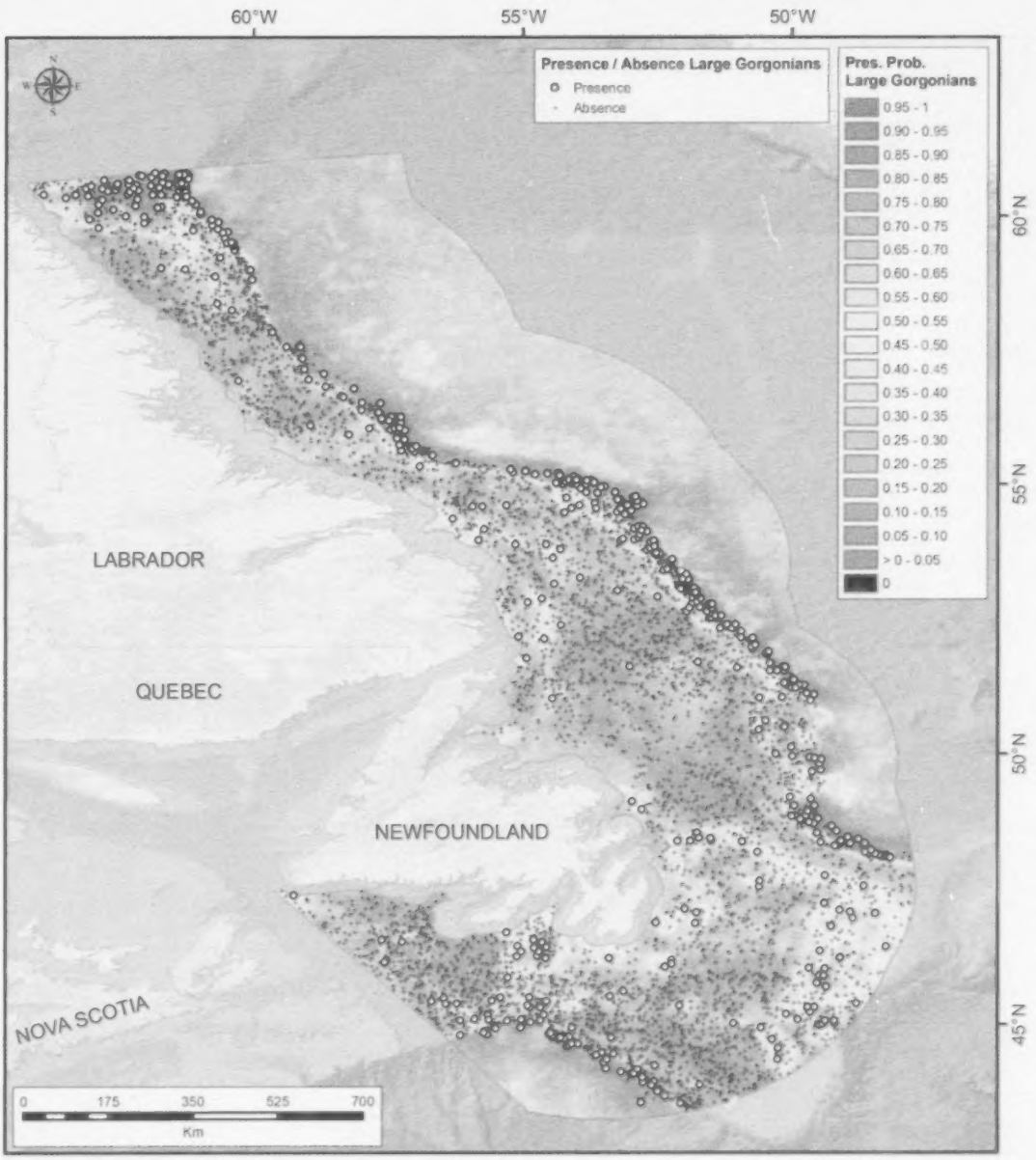

Figure 35. Presence and absence observations and predictions of presence probability of the optimal random forest model of large gorgonian coral presence and absence data collected from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

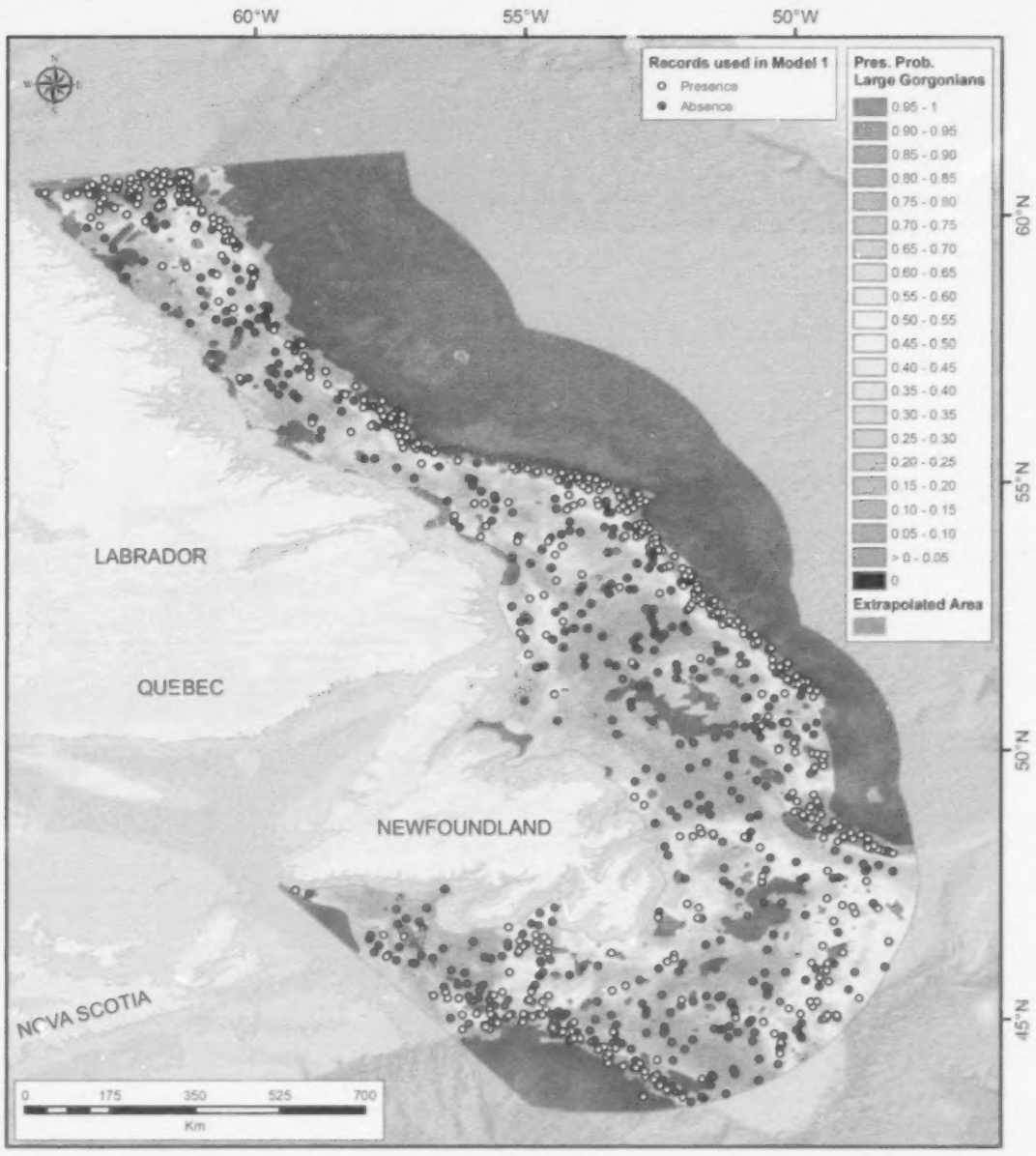

Figure 36. Map of the 1028 data observations (514 presences and 514 absences) of large gorgonian corals used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of large gorgonian corals generated from Model 1 and areas of model extrapolation.

Of all 66 environmental predictor variables used in the model, Bottom Salinity Average Range was the most important for the classification of the large gorgonian coral presence and absence data (Figure 37). Prior to spatial interpolation, this variable displayed a right-skewed distribution (Guijarro et al., in prep.). Examination of the Q-Q plot revealed a strong spatial pattern to those

data points over- and under-predicted by a normal distribution, with over-predicted points located mainly in the deep waters beyond the Labrador Shelf, and under-predicted points located along the Newfoundland and Labrador Slopes. Bottom Salinity Average Range was followed closely by Depth, Bottom Temperature Average Range, Bottom Salinity Average Minimum and Slope. Partial dependence plots for the top 6 predictor variables are shown in Figure 38. Along the Bottom Salinity Average Range gradient, the highest predicted presence probabilities occurred between 0 and 0.2. Values in this range coincided with both over- and under-predicted values near the shelf break and in the deep waters beyond. Most of this area is considered extrapolated by the model. The fit between predicted and observed values in the kriging model was relatively good, with only slight over-prediction of values between 0 and 0.2. Some points could therefore be predicted higher than their true values and slightly outside the range of highest presence probability identified in the partial plot (Figure 38). Along the Depth gradient, presence probability increased gradually beginning at ~ 400 m and then decreased slightly prior to 1500 m.

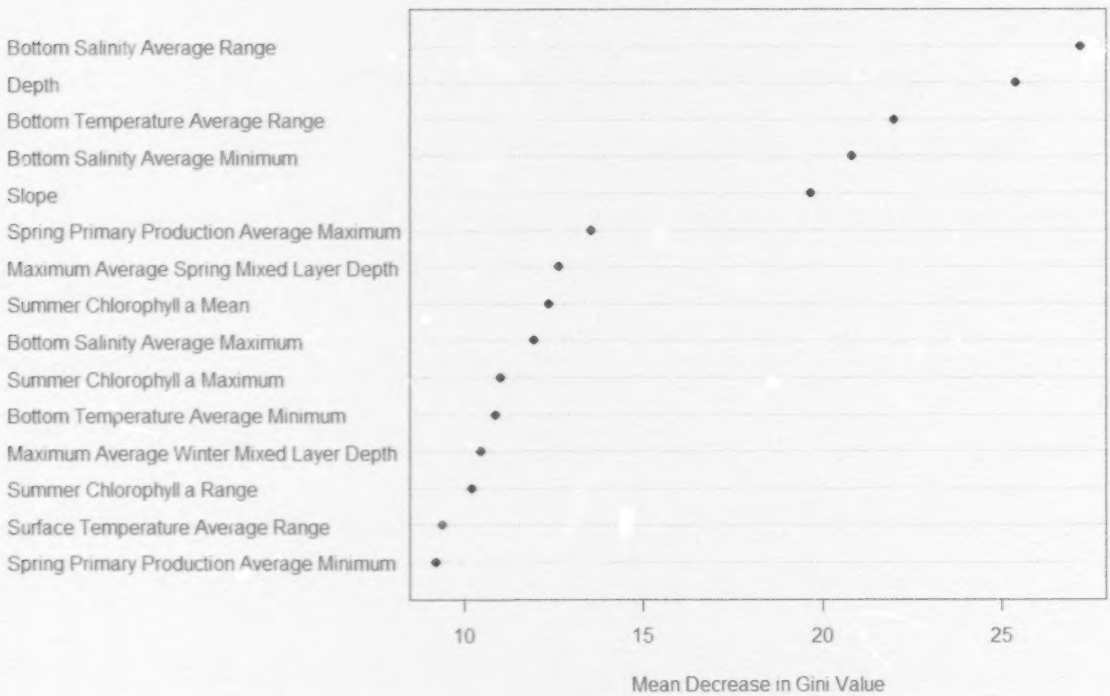

Figure 37. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting large gorgonian coral presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini value the more important the variable is for predicting the response data.

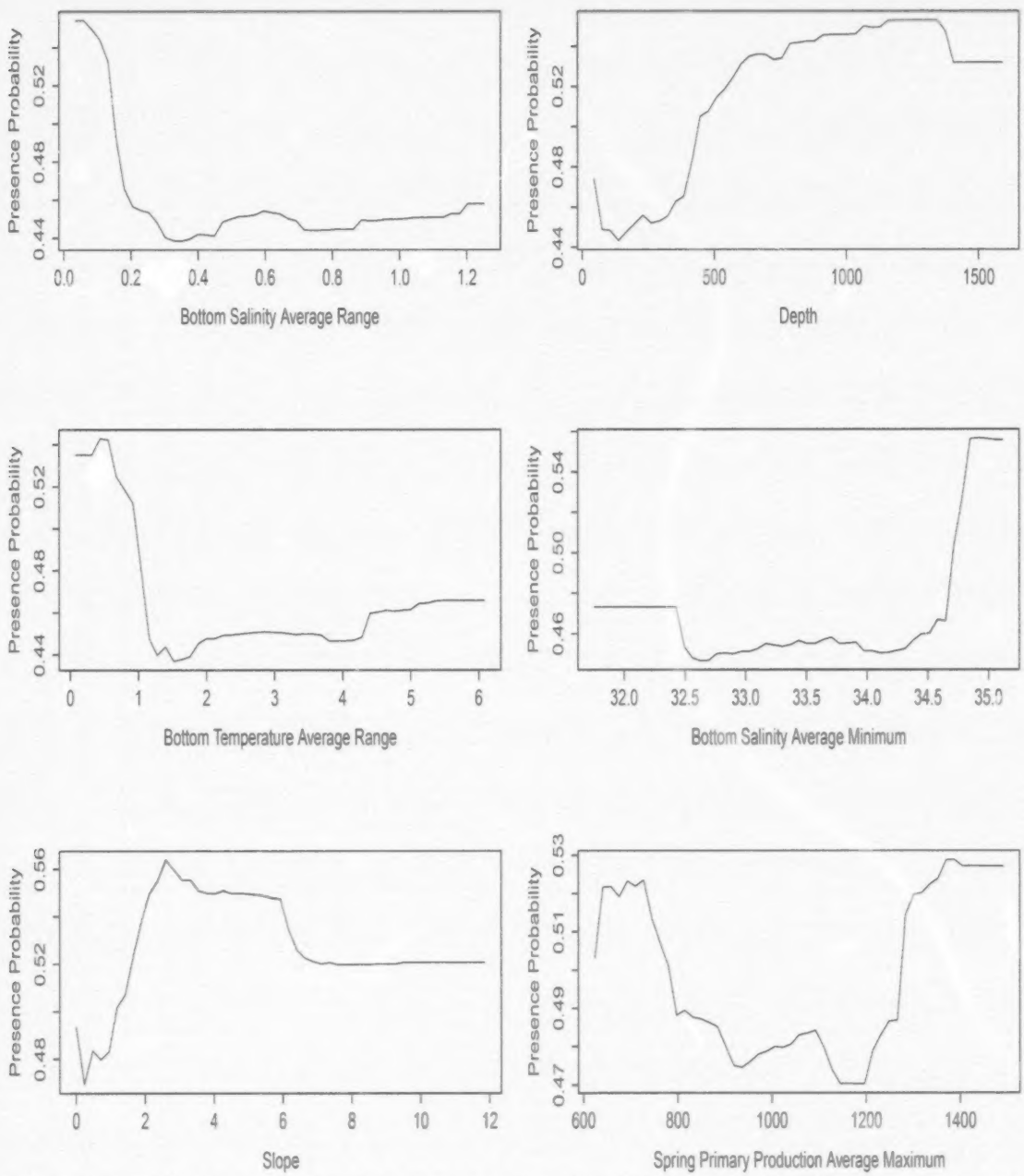

Figure 38. Partial dependence plots of the top 6 predictors from the optimal random forest model of large gorgonian coral presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 13 shows the accuracy measures for the random forest model on all large gorgonian coral presence absence data (514 presences 5651 absences; Model 2) and a threshold equal to species prevalence (0.08). The average AUC calculated from this model was slightly lower than that of Model 1 (0.806 compared to 0.811 of Model 1). Sensitivity and specificity were also slightly lower than Model 1.

The surface of predicted presence probability of large gorgonian corals generated from Model 2 is presented in Figure 39. The areas of high presence probability along the slopes were much reduced in this model. The highest predicted presence probability of large gorgonian corals occurred on the northeast slope off Saglek Bank where there was a large concentration of presence observations (Figure 40). Figure 41 depicts the classification of small gorgonian presence probability into presence and absence categories based on the prevalence threshold of 0.08. In this map, all presence probability values generated from Model 2 that were greater than 0.08 were classified as presence, while values less than 0.08 were classed as absence. The slopes of Newfoundland and Labrador are predicted as presence of large gorgonian corals. A number of small pockets of coral presence were scattered across the continental shelf.

Table 13. Accuracy measures and confusion matrix from 10-fold cross validation from random forest model of presence and absence of large gorgonian corals collected within the Newfoundland and Labrador Region between 2003 and 2015. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.852		Absence	Presence				
2	0.794	Absence	4330	1321	5651	0.234	0.726	0.766
3	0.784	Presence	141	373	514	0.274		
4	0.859							
5	0.818							
6	0.761							
7	0.748							
8	0.784							
9	0.854							
10	0.808							
Mean	**0.806**							
SD	**0.039**							

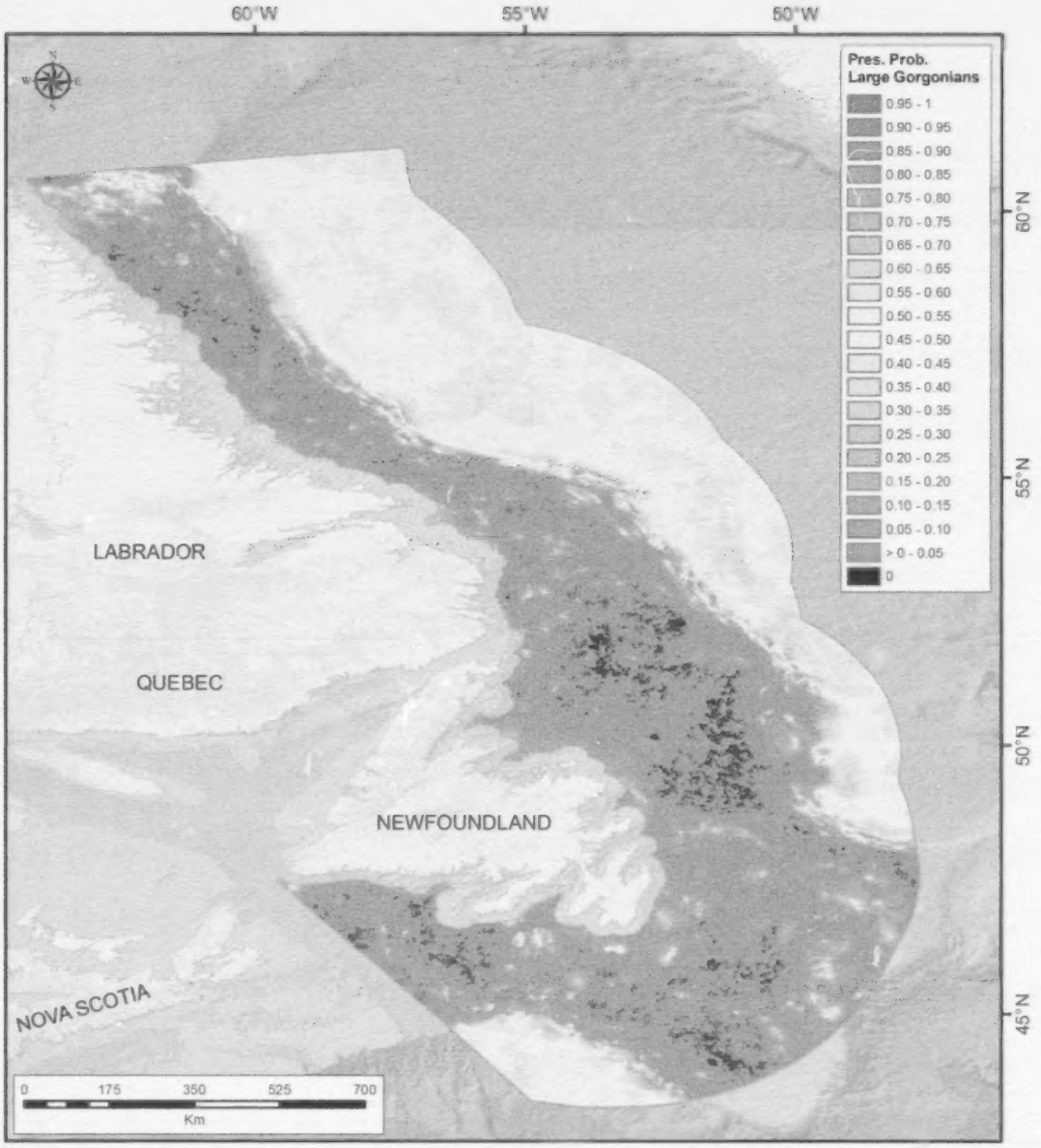

Figure 39. Predictions of presence probability from the unbalanced random forest model of large gorgonian coral presence and absence data collected from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

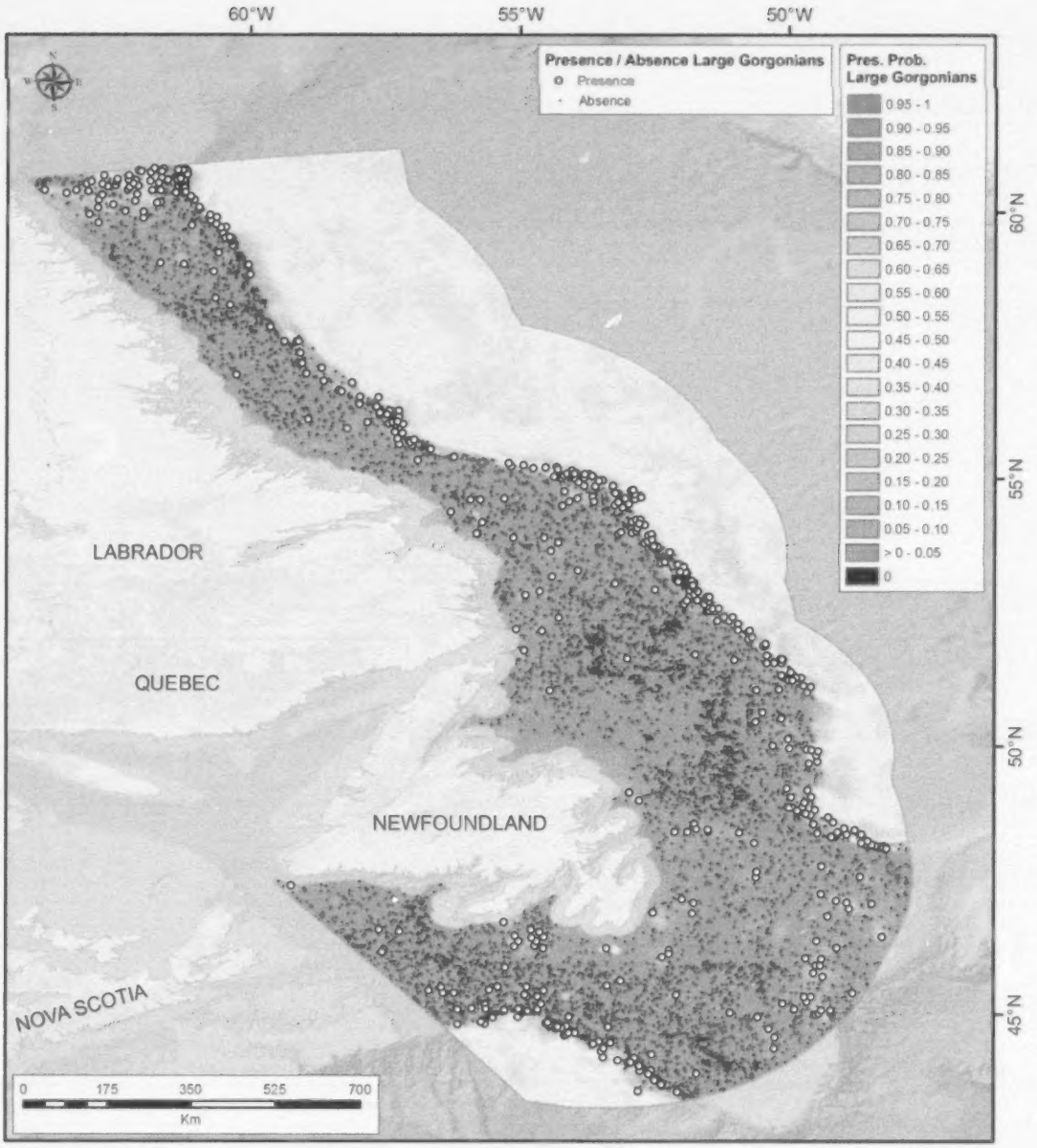

Figure 40. Presence and absence observations and predictions of presence probability from the unbalanced random forest model of large gorgonian coral presence and absence data collected from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

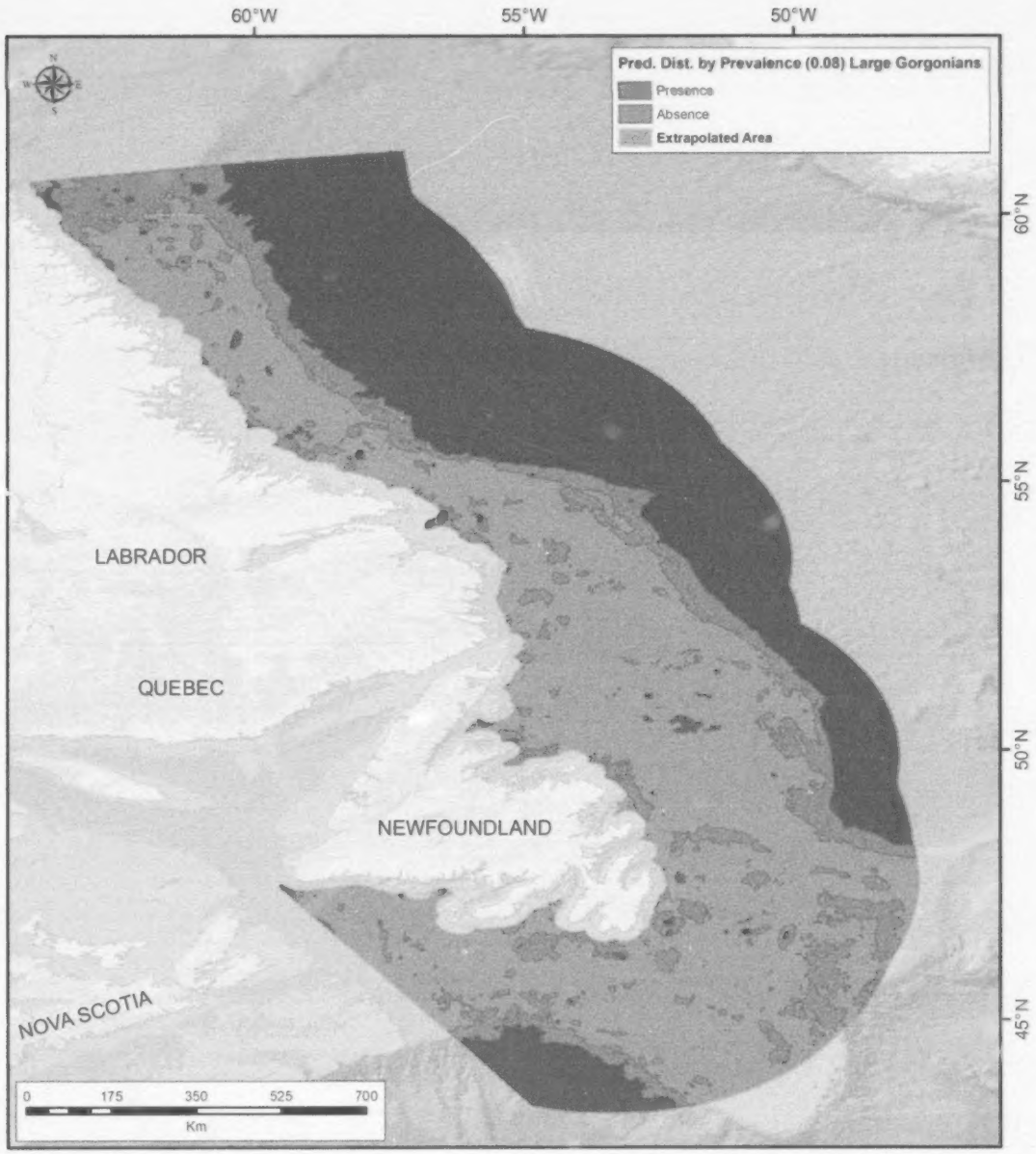

Figure 41. Predicted distribution (Pred. Dist.) of large gorgonian corals in the Newfoundland and Labrador Region based on the prevalence threshold of 0.08 of large gorgonian presence and absence data used in Model 2. Also shown are the areas of model extrapolation. Grey polygon may appear dark red or blue when overlain on the presence-absence surface.

The importance of the environmental predictor variables for predicting the presence probability of the large gorgonian coral catch data is presented in Figure 42. In this model, Depth (a non-interpolated variable) was the most important variable for the classification of the large

gorgonian coral presence and absence data. This variable was followed by Bottom Temperature Average Range, Bottom Salinity Average Range, Bottom Salinity Average Minimum Average and Slope. Partial dependence of the large gorgonian coral presence and absence data on the top 6 predictor variables is shown in Figure 43. Along the Depth gradient, presence probability of large gorgonian corals increased rapidly at ~ 500 m and plateaued at ~1000 m. Along the gradient in Bottom Temperature Average Range, the highest presence probability occurred at ~ 0.5 °C.

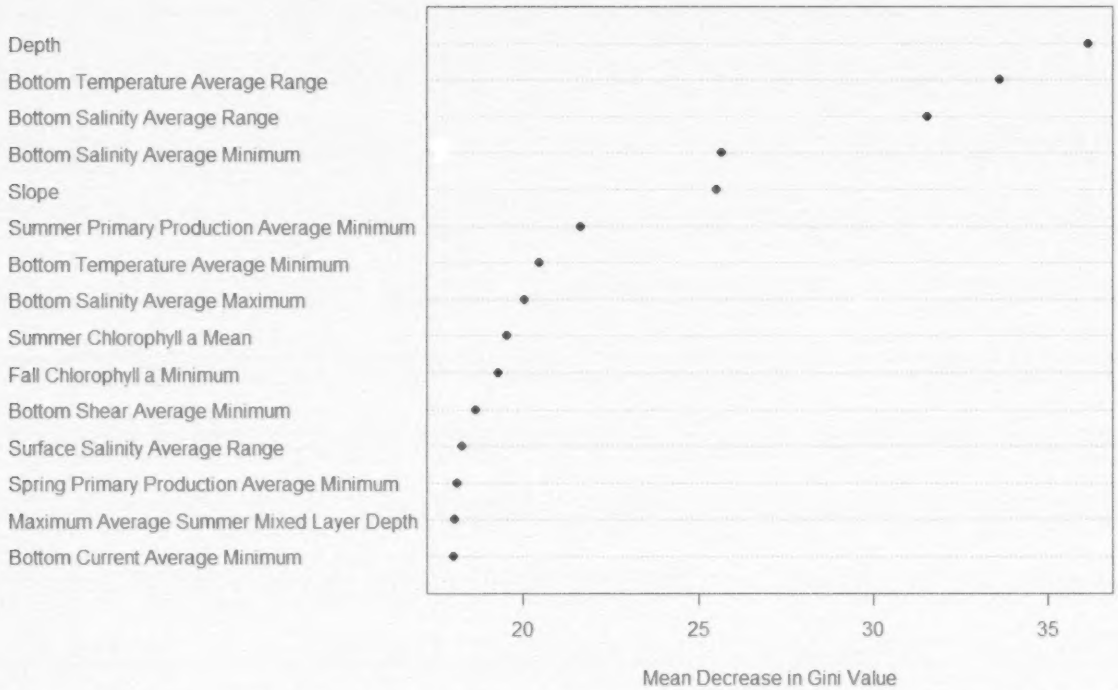

Figure 42. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the unbalanced random forest model of large gorgonian coral presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini Value the more important the variable is for predicting the response data.

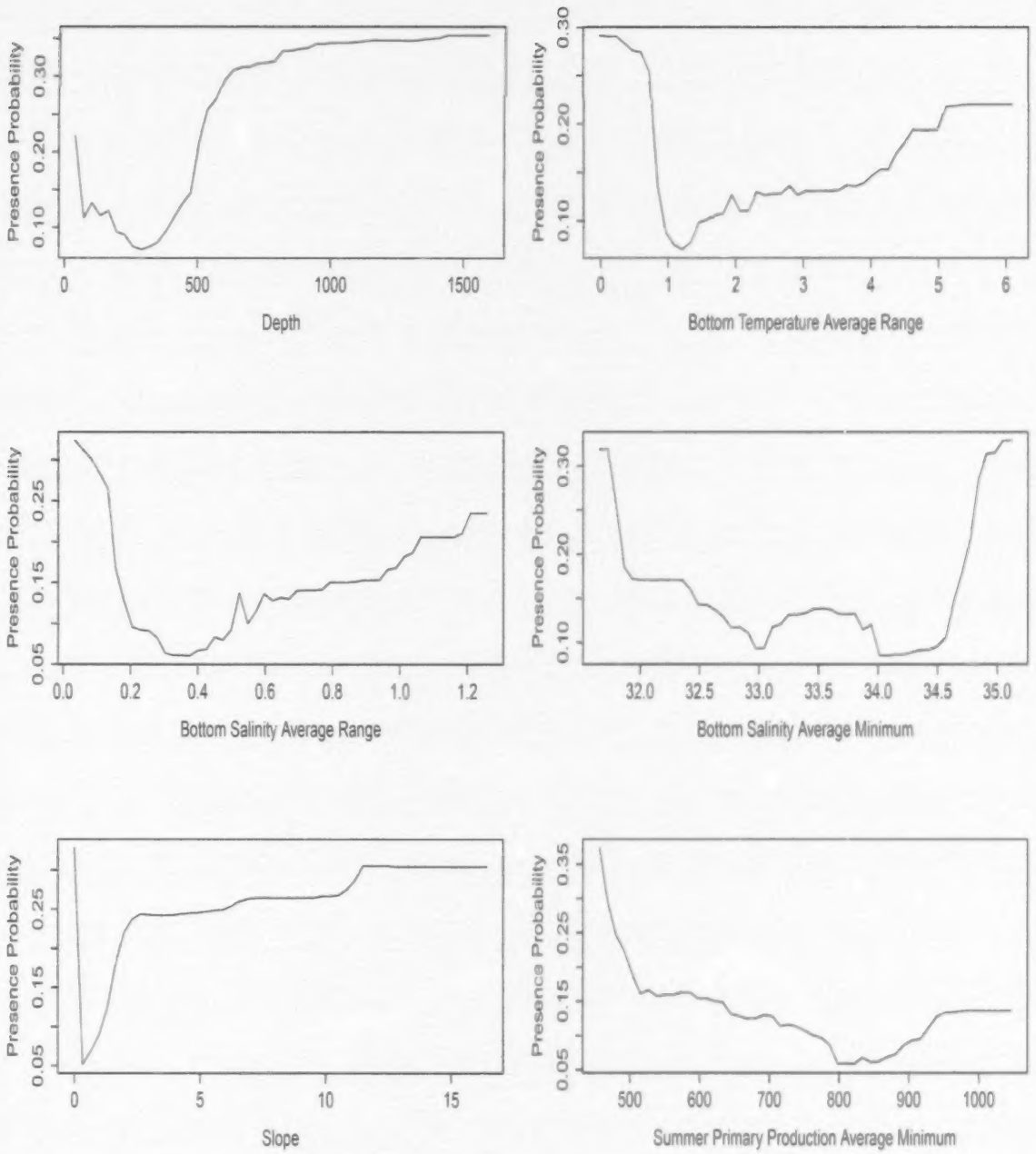

Figure 43. Partial dependence plots of the top 6 predictors from the unbalanced random forest model of large gorgonian coral presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model Selection

The random forest model using all available large gorgonian records and an unbalanced species prevalence and threshold equal to 0.08 (Model 2) was chosen as the best predictor of large gorgonian coral distribution in the Newfoundland and Labrador Region. Model 1 (balanced species prevalence) was considered a poor predictor of presence probability of large gorgonians due to its exaggeration of high presence probability beyond the location of presence data, particularly slopes off Labrador and the Northeast Newfoundland Shelf. This phenomenon is likely due to random down-sampling of the absence data.

Prediction of Large Gorgonian Coral Biomass Using Random Forest

The accuracy measures of the regression random forest model on mean large gorgonian coral biomass per grid cell are presented in Table 14. The highest R^2 value was 0.690, while the average was 0.203 ± 0.218 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.017 ± 0.012 SD. This model explained a low percentage of variance in the biomass data (average = 5.70% ± 3.34 SD).

Table 14. Accuracy measures from 10-fold cross validation of random forest model of mean large gorgonian coral biomass (kg) per grid cell recorded from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.250	3.793	0.013	5.63
2	0.195	6.593	0.023	5.79
3	8.210×10^{-5}	3.979	0.014	6.03
4	0.022	4.499	0.016	1.59
5	0.213	13.348	0.046	9.90
6	0.049	6.406	0.022	7.65
7	0.690	0.837	0.003	2.77
8	0.448	3.004	0.010	0.48
9	0.096	4.616	0.016	10.74
10	0.065	0.850	0.003	6.39
Mean	**0.203**	**4.792**	**0.017**	**5.70**
SD	**0.218**	**3.578**	**0.012**	**3.34**

Figures 44 and 45 show the predicted biomass surface of large gorgonian corals. The majority of the spatial extent was predicted to have low (> 0 – 0.69 kg) large gorgonian biomass, even in areas where (> 0 – 25.09 kg) large gorgonian corals catch were recorded. The highest biomass prediction (reaching up to 175.14 kg) occurred on the slope off Saglek Bank. This area of high biomass was associated with a cluster of large biomass values (Figure 45).

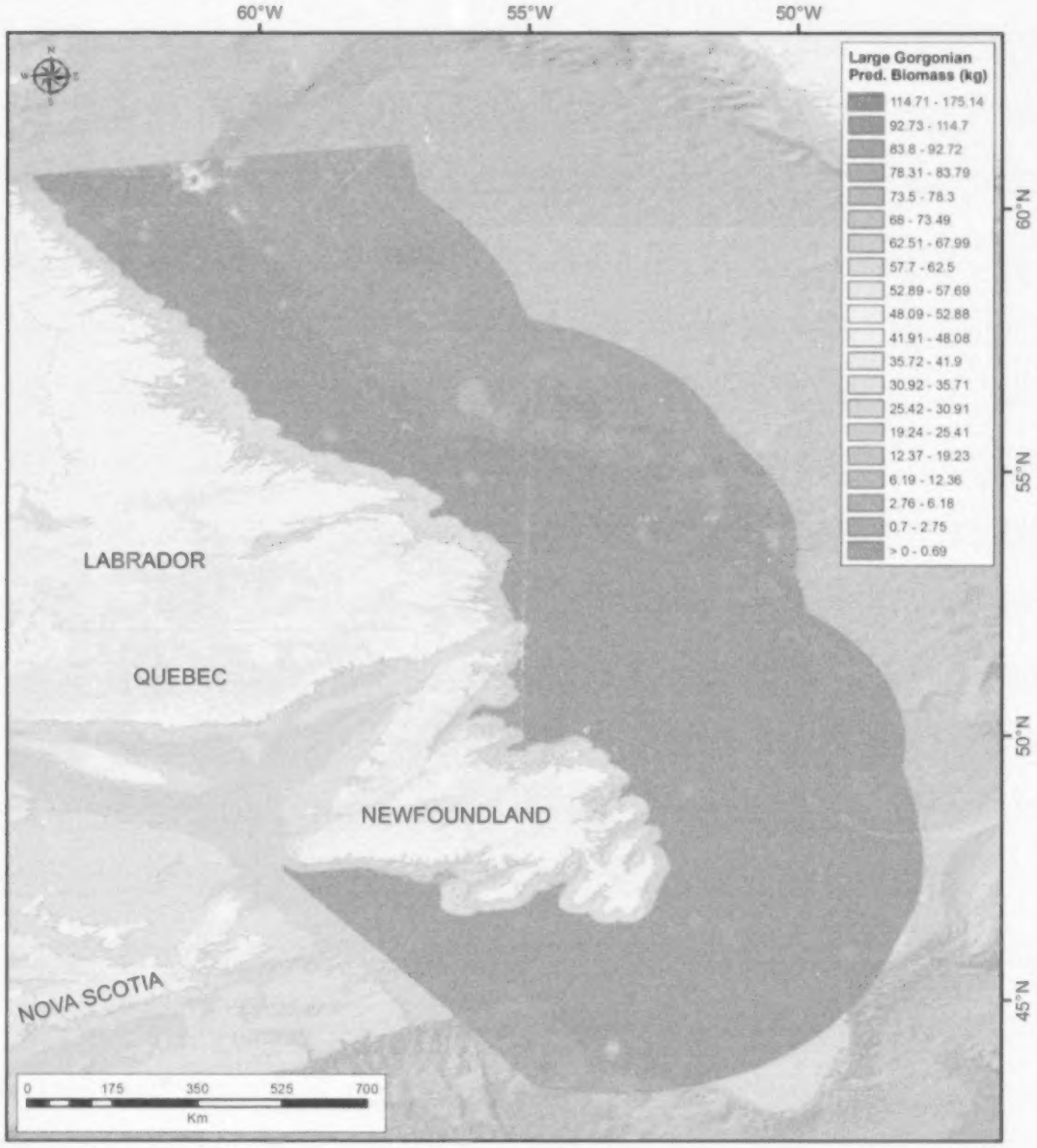

Figure 44. Predictions of biomass (kg) of large gorgonian corals from catch recorded in DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

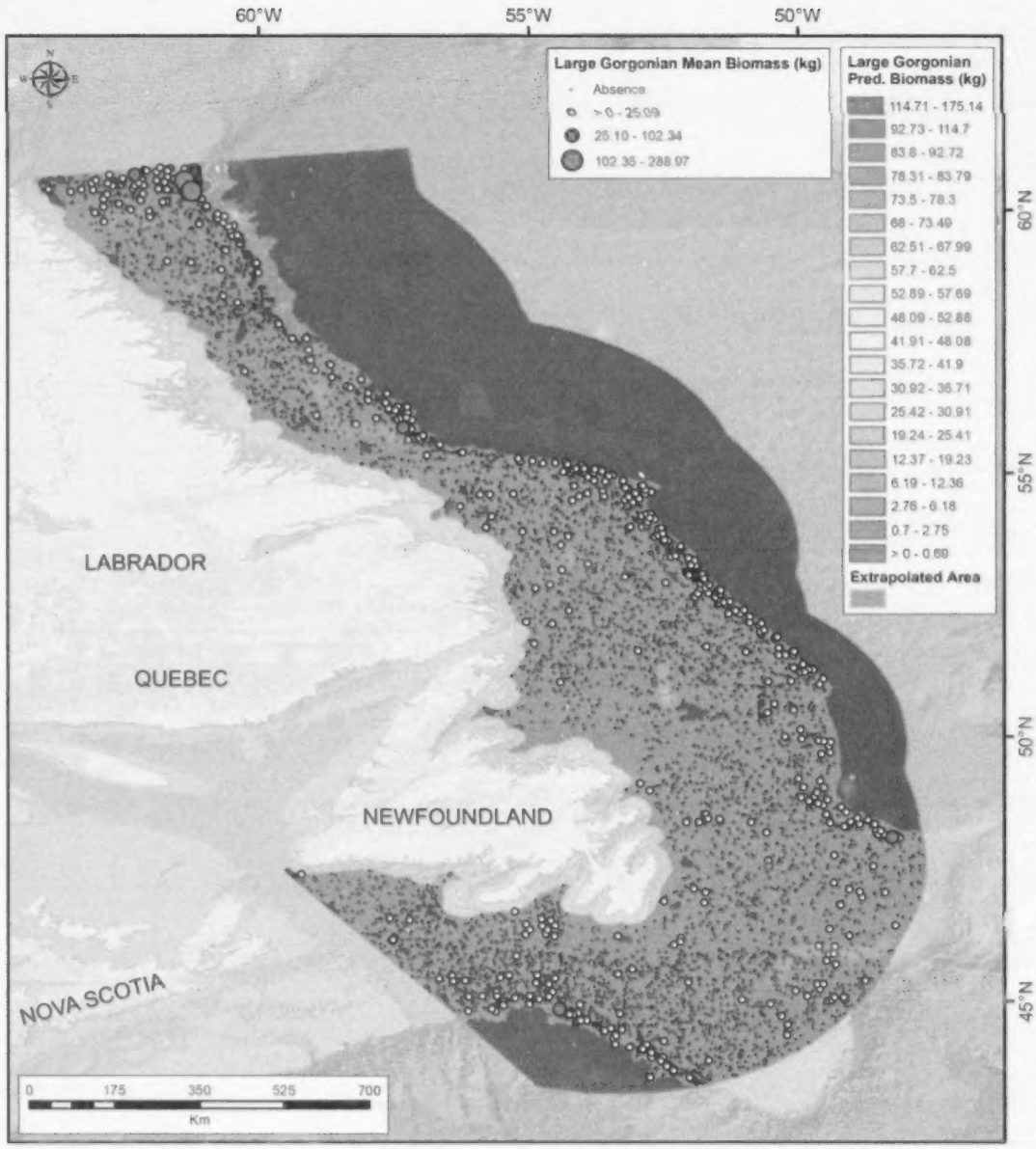

Figure 45. Predictions of biomass (kg) of large gorgonian corals from catch data recorded in DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting large gorgonian coral biomass are shown in Figure 46. Summer Primary Production Average Minimum was the most important variable in the model. This variable displayed a near-normal distribution prior to spatial interpolation (Guijarro et al., in prep.). Examination of the Q-Q plot revealed a spatial pattern to

those data points over- and under-predicted by a normal distribution, with over-predicted points located mainly in the northern portion of the study extent on Nain and Saglek Banks, and under-predicted points located on the Grand Banks and the Northeast Newfoundland Shelf. Summer Primary Production Average Minimum was followed very distantly by Fall Chlorophyll a Minimum and the remaining variables in the model. The partial dependence of large gorgonian coral biomass on the top 6 most important variables is shown in Figure 47. Predicted biomass was the highest at the lowest Summer Primary Production Average Minimum values (< 500 mg C m^{-2} day^{-1}). Values in this range coincided with those data points under-predicted by a normal distribution. The fit between predicted and observed values in the kriging model was good, with slight over- prediction of data points in that range. Some points could therefore be predicted higher than their true values and slightly outside the range of highest predicted biomass identified in the partial plot.

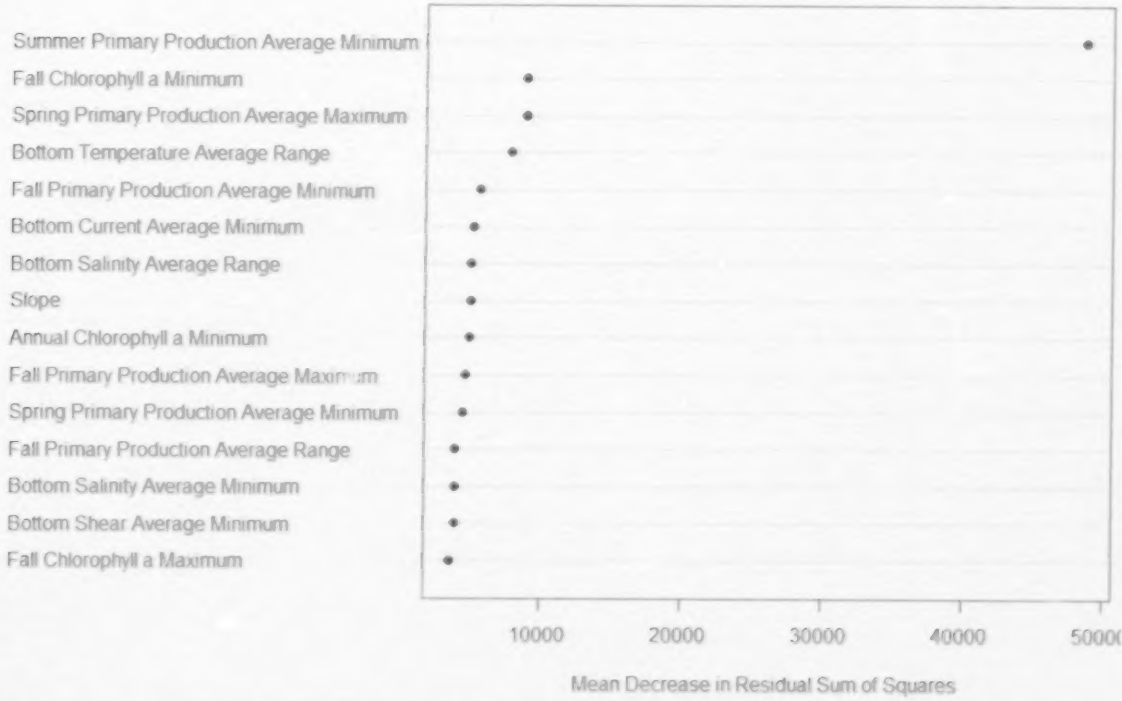

Figure 46. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on mean large gorgonian coral biomass per grid cell. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

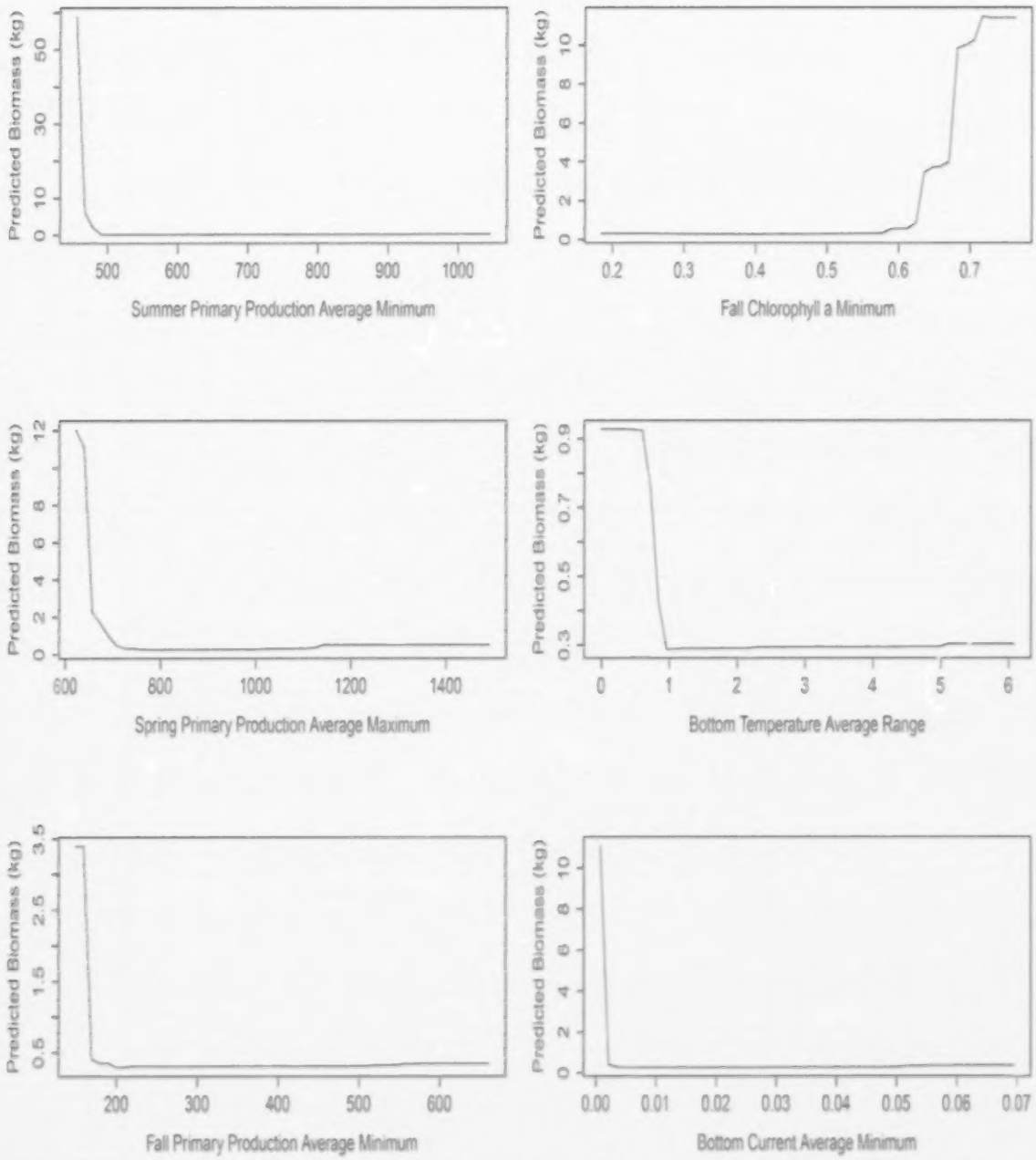
Predicted Biomass (kg)
0 10 30 50
500 600 700 800 900 1000
Summer Primary Production Average Minimum
Predicted Biomass (kg)
0 2 4 6 8 10
0.2 0.3 0.4 0.5 0.6 0.7
Fall Chlorophyll a Minimum
Predicted Biomass (kg)
0 2 4 6 8 12
600 800 1000 1200 1400
Spring Primary Production Average Maximum
Predicted Biomass (kg)
0.3 0.5 0.7 0.9
0 1 2 3 4 5 6
Bottom Temperature Average Range
Predicted Biomass (kg)
0.5 1.5 2.5 3.5
200 300 400 500 600
Fall Primary Production Average Minimum
Predicted Biomass (kg)
0 2 4 6 8 10
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07
Bottom Current Average Minimum

Figure 47. Partial dependence plots of the top six predictors from the random forest model of large gorgonian biomass data collected within the Newfoundland and Labrador Region between 2003 and 2015, ordered left to right from the top. Predicted biomass is shown on the *y*-axis of each graph.

Small Gorgonian Corals

Data Sources and Distribution

Small gorgonian coral catch data was collected over a span of 13 years from 2003 to 2015 and consisted of 370 presence and 4967 absence records after the data were grouped to 1 record per cell (Table 15). Note that although Spanish records of small gorgonians exist, they were not selected for use in presence-absence models when the data were randomly grouped to 1 record per cell (but were used in the calculation of mean biomass per grid cell for biomass random forest models below). Absence records were distributed relatively evenly across the study extent (Figure 48). However, presence records had a highly uneven distribution and were concentrated mainly along the slopes of Newfoundland and Labrador, although some records occurred on the shelf and banks. The highest mean biomass records (up to 2.80 kg) were located on the slope southwest of Grand Bank in the NAFO 3O closure area.

Table 15. Number of presence and absence records of small gorgonian coral catch recorded from DFO multispecies surveys, DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

Year	Total of number of presences	Total of number of absences
2003	8	74
2004	24	251
2005	22	362
2006	40	395
2007	46	376
2008	51	402
2009	33	483
2010	44	605
2011	19	515
2012	26	563
2013	34	389
2014	19	428
2015	4	124

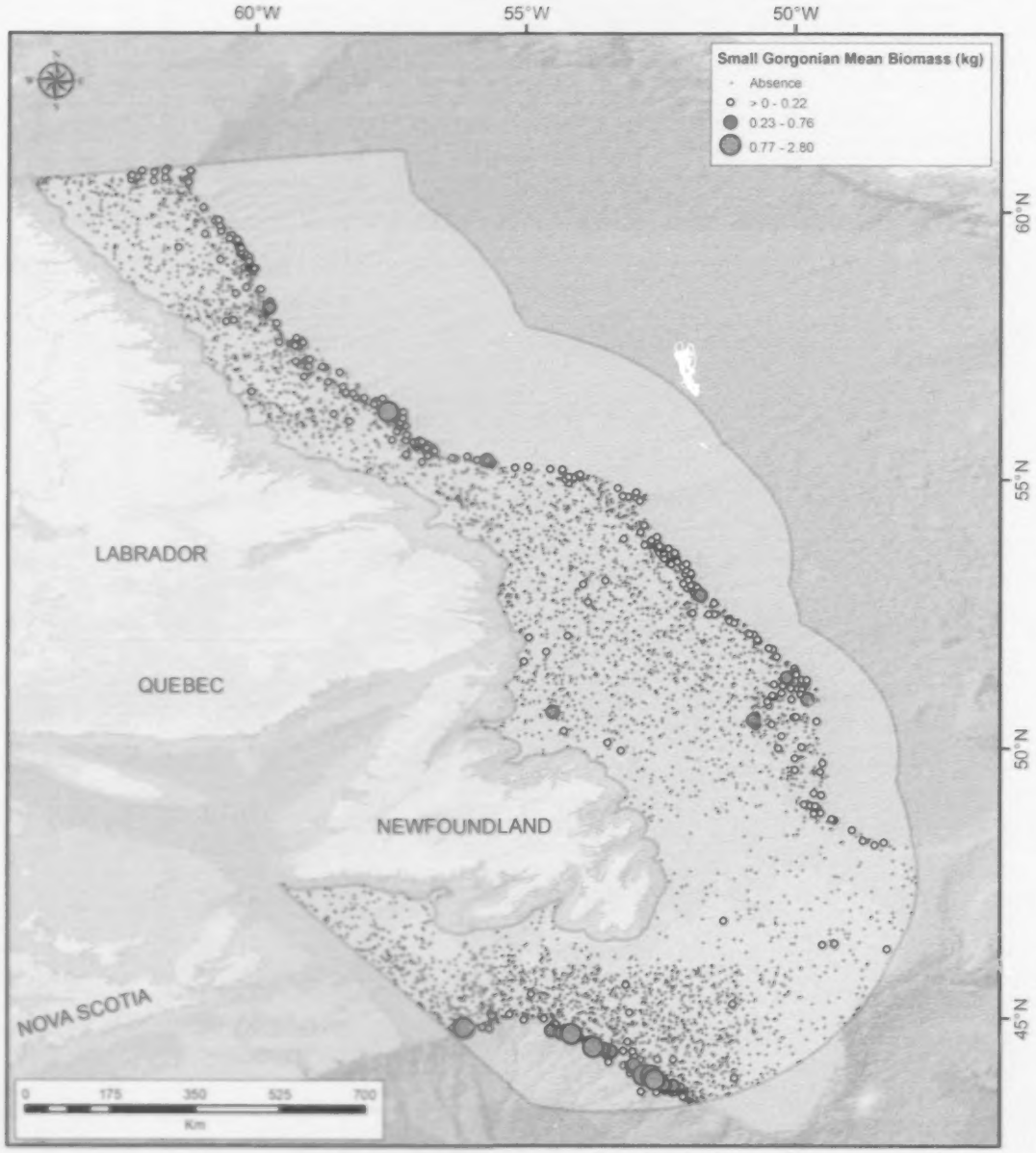

Figure 48. Mean biomass (kg) per grid cell of small gorgonian corals recorded from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

Model 1 – Balanced Species Prevalence

Accuracy measures (mean AUC, sensitivity, and specificity) for the random forest model on balanced species prevalence (370 presences and 370 absences; Model 1) are presented in Table 16. The average AUC was 0.864, indicating very good model performance. The highest mean AUC of 0.884 was associated with Model Run 8 and is therefore considered the optimal model for the prediction of the small gorgonian corals response data. The sensitivity and specificity measures of this model were 0.835 and 0.832, respectively. The confusion matrix of the optimal model is also presented in Table 16. Class error for both the presence and absence classes were relatively low (0.165 and 0.168, respectively).

Table 16. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of small gorgonian corals within the Newfoundland and Labrador Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 8) which is considered the optimal model for predicting the presence probability of small gorgonian corals.

Model Run	AUC	Sensitivity	Specificity
1	0.876	0.808	0.800
2	0.861	0.822	0.816
3	0.873	0.835	0.803
4	0.880	0.830	0.838
5	0.846	0.808	0.765
6	0.850	0.819	0.778
7	0.868	0.805	0.822
8	**0.884**	**0.835**	**0.832**
9	0.851	0.814	0.762
10	0.855	0.816	0.805
Mean	**0.864**	**0.818**	**0.802**
SD	**0.014**	**0.011**	**0.027**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	308	62	370	0.168
Presence	61	309	370	0.165

The presence probability prediction surface of the small gorgonian corals is presented in Figure 49. The highest predictions of presence probability occurred along the slopes of Newfoundland and Labrador. These areas corresponded well with the spatial distribution of presence records (see Figure 50). However, the model appears to moderately extrapolate areas of presence probability beyond the location of presence observations.

The actual presence and absence data observations (370 presences and 370 absences) used in the optimal run of Model 1 showed some slight spatial bias across the study area (Figure 51). Few absence points were selected from the slopes. Also shown in this figure are the areas of model extrapolation. Deep water beyond the slope was considered extrapolated area. Smaller pockets of extrapolated area were distributed across the shelf, particularly along the coast of Labrador.

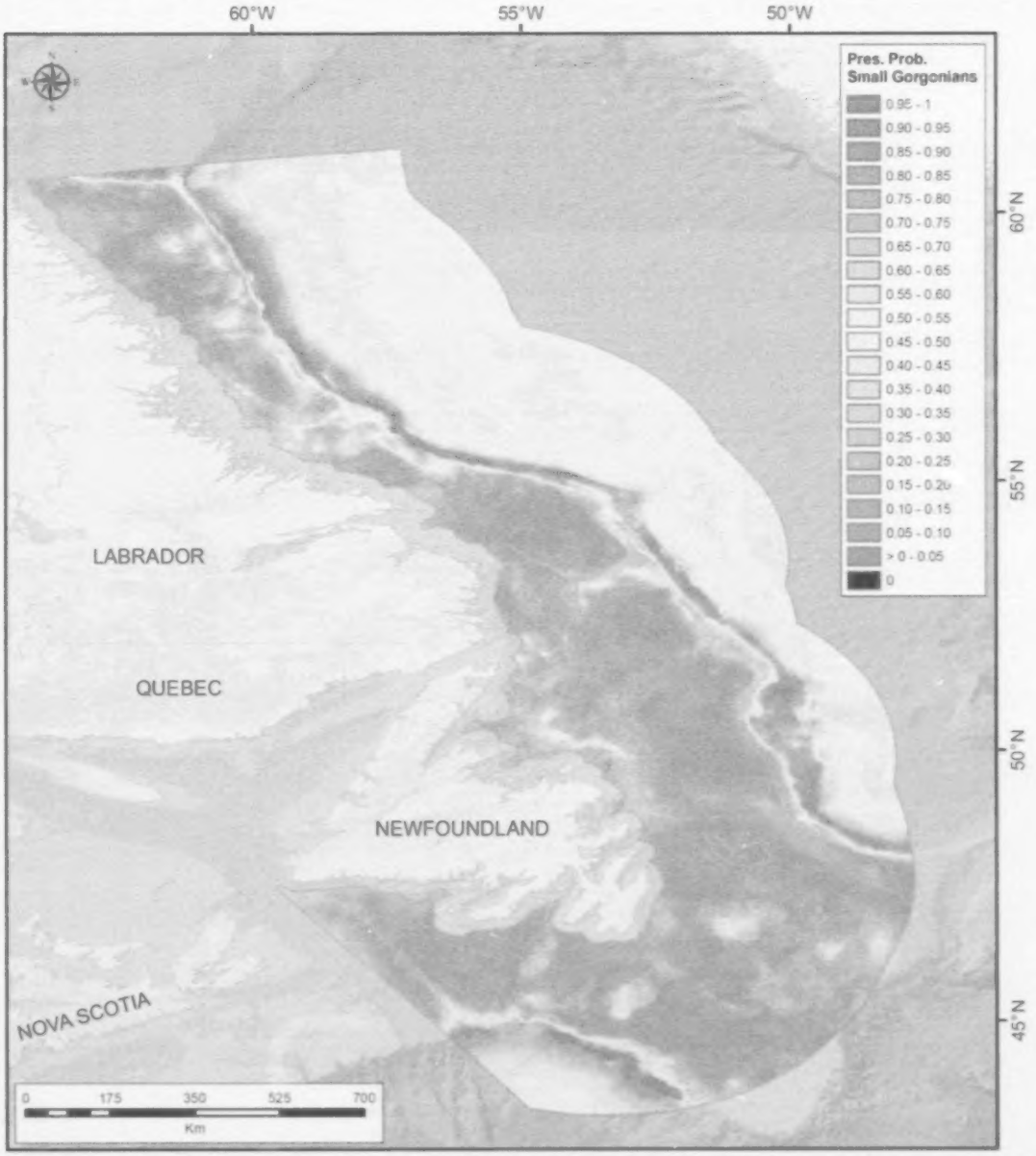

Figure 49. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of small gorgonian coral presence and absence data collected from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

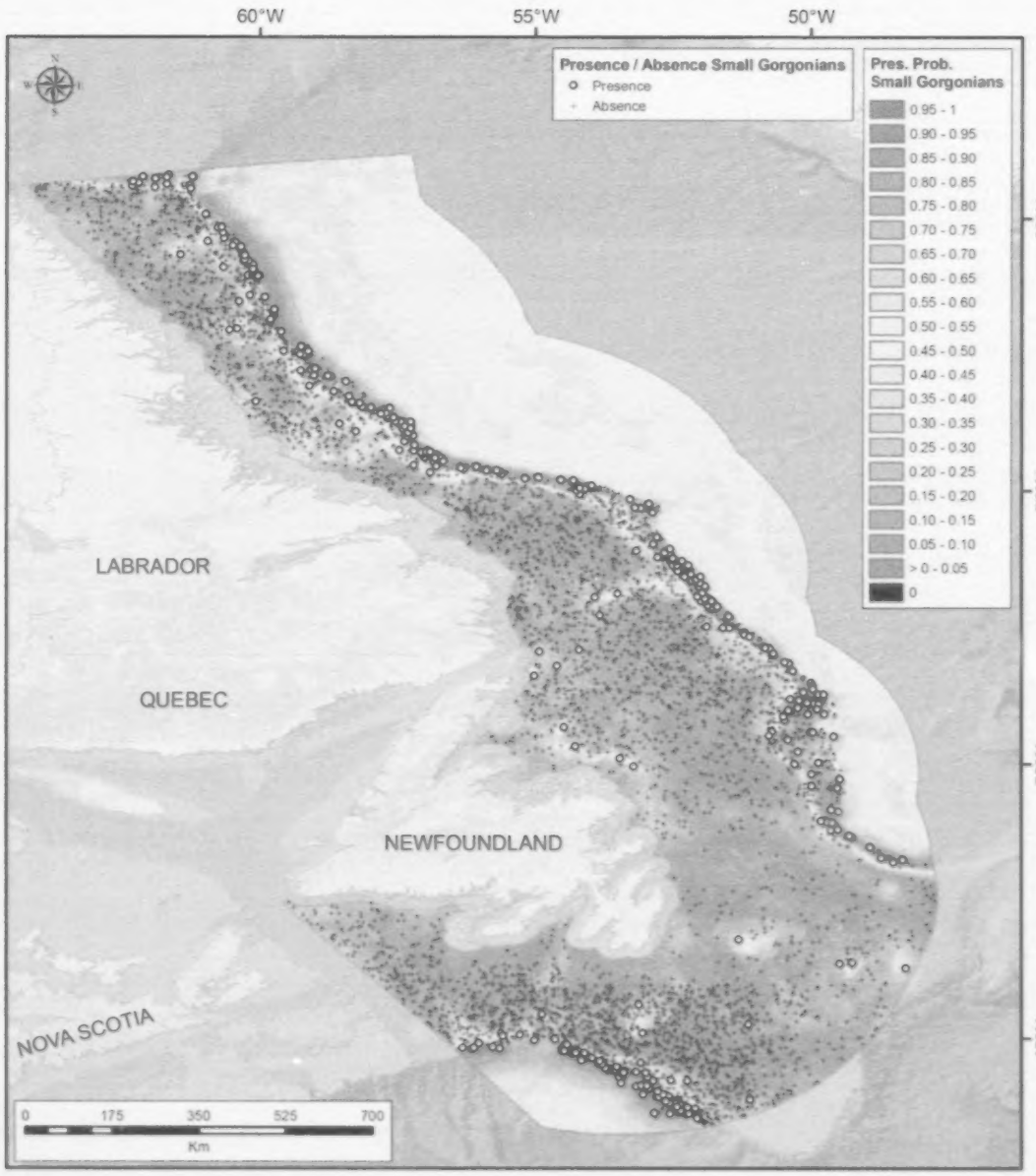

Figure 50. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of small gorgonian corals presence and absence data collected from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

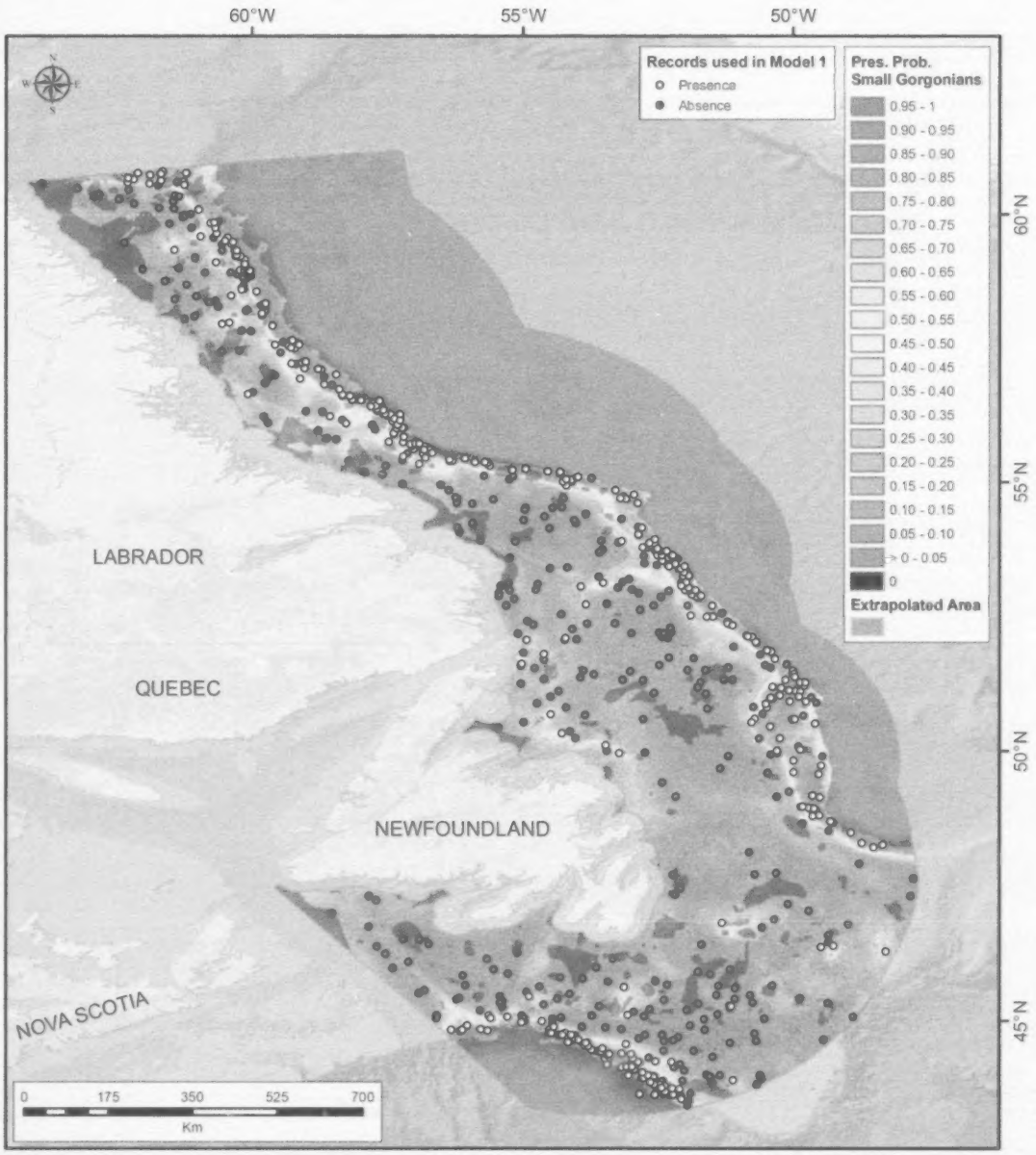

Figure 51. Map of the 740 data observations (370 presences and 370 absences) of small gorgonian corals used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of small gorgonian corals generated from Model 1.

Of all 66 environmental predictor variables used in the model, Slope (non-interpolated variable) was the most important for the classification of the small gorgonian coral presence and absence data (Figure 52). This variable was followed closed by Bottom Temperature Average Minimum, Depth, and Bottom Salinity Average Minimum. Partial dependence plots for the top 6 predictor

variables are shown in Figure 53. Presence probability of small gorgonians rapidly increased at 2° along the Slope gradient, and at 3°C along the gradient in Bottom Temperature Average Minimum.

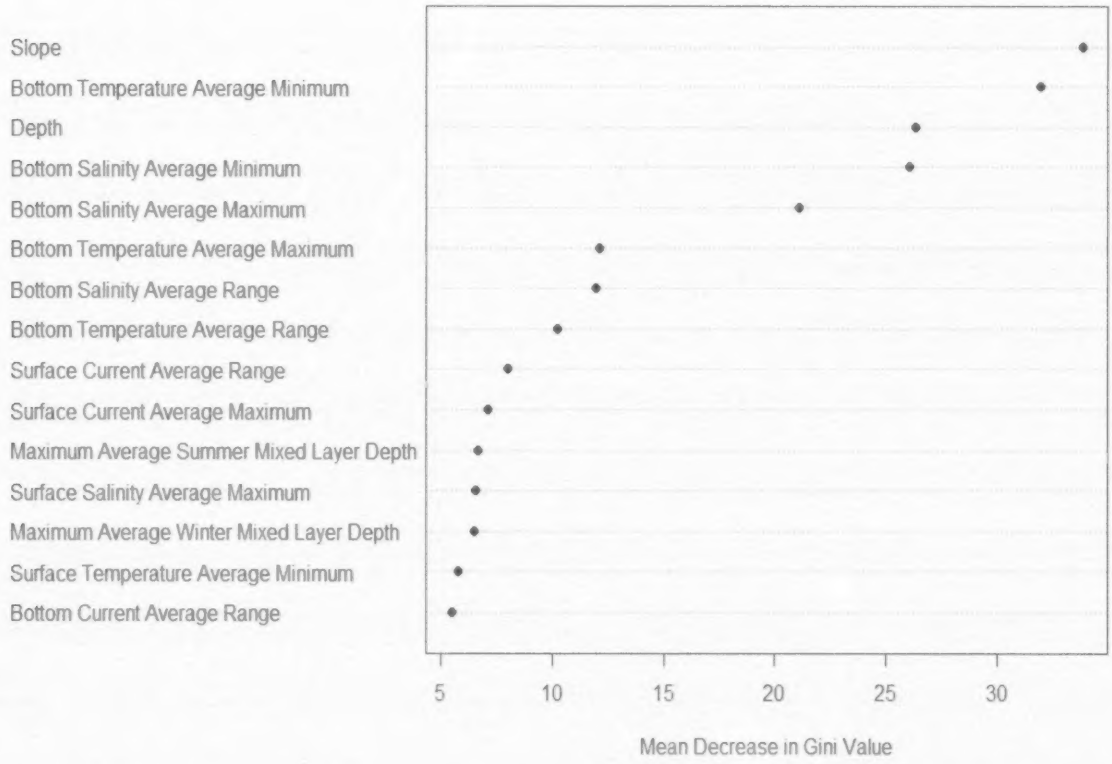

Figure 52. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting small gorgonian coral presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini value the more important the variable is for predicting the response data.

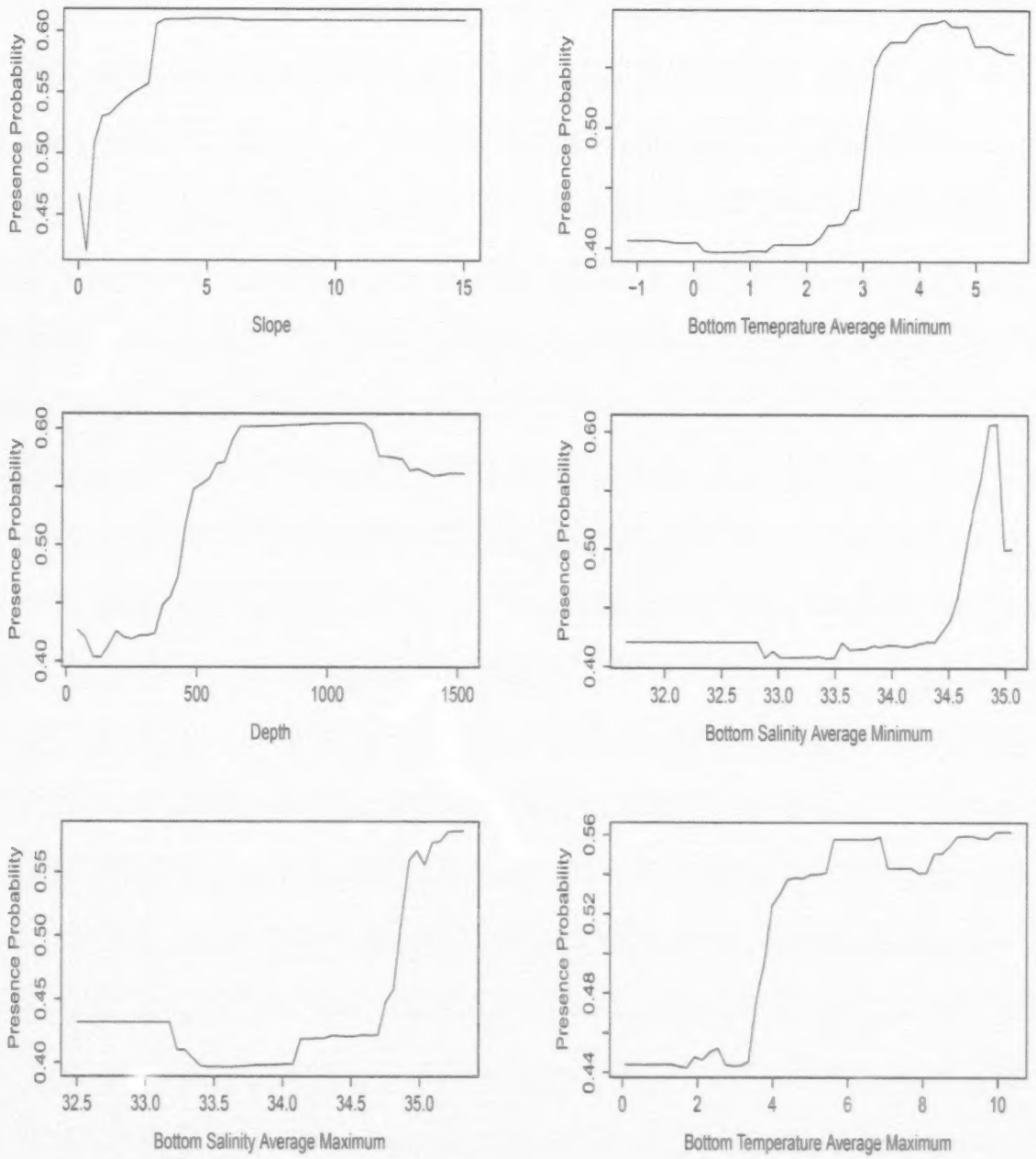

Figure 53. Partial dependence plots of the top 6 predictors from the optimal random forest model of small gorgonian coral presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 17 shows the accuracy measures for the random forest model on all small gorgonian coral presence and absence data (370 presences and 4967 absences; Model 2) and a threshold equal to species prevalence (0.07). The average AUC calculated from this model was slightly lower than that of Model 1 (0.859 compared to 0.864 of Model 1). Sensitivity and specificity measures of Model 2 were also lower than that of Model 1.

The surface of predicted presence probability of small gorgonian corals generated from Model 2 is presented in Figure 55. The areas of high presence probability of small gorgonians from Model 1 are much reduced in this model. The slope southwest of Grand Bank had the highest presence probability of small gorgonians. These areas of high presence probability corresponded well with the spatial distribution of presence records (Figure 55). Figure 56 depicts the classification of small gorgonian presence probability into presence and absence categories based on the prevalence threshold of 0.07. In this map, all presence probability values generated from Model 2 that were greater than 0.07 were classified as presence, while values less than 0.07 were classed as absence. The slopes and small pockets across the shelf were predicted as presence of small gorgonians.

Table 17. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model presence and absence of small gorgonian corals within the Newfoundland and Labrador Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.842		**Absence**	**Presence**				
2	0.881	**Absence**	3975	992	4967	0.200	0.800	0.800
3	0.814	**Presence**	74	296	370	0.200		
4	0.801							
5	0.887							
6	0.887							
7	0.918							
8	0.840							
9	0.900							
10	0.818							
Mean	**0.859**							
SD	**0.041**							

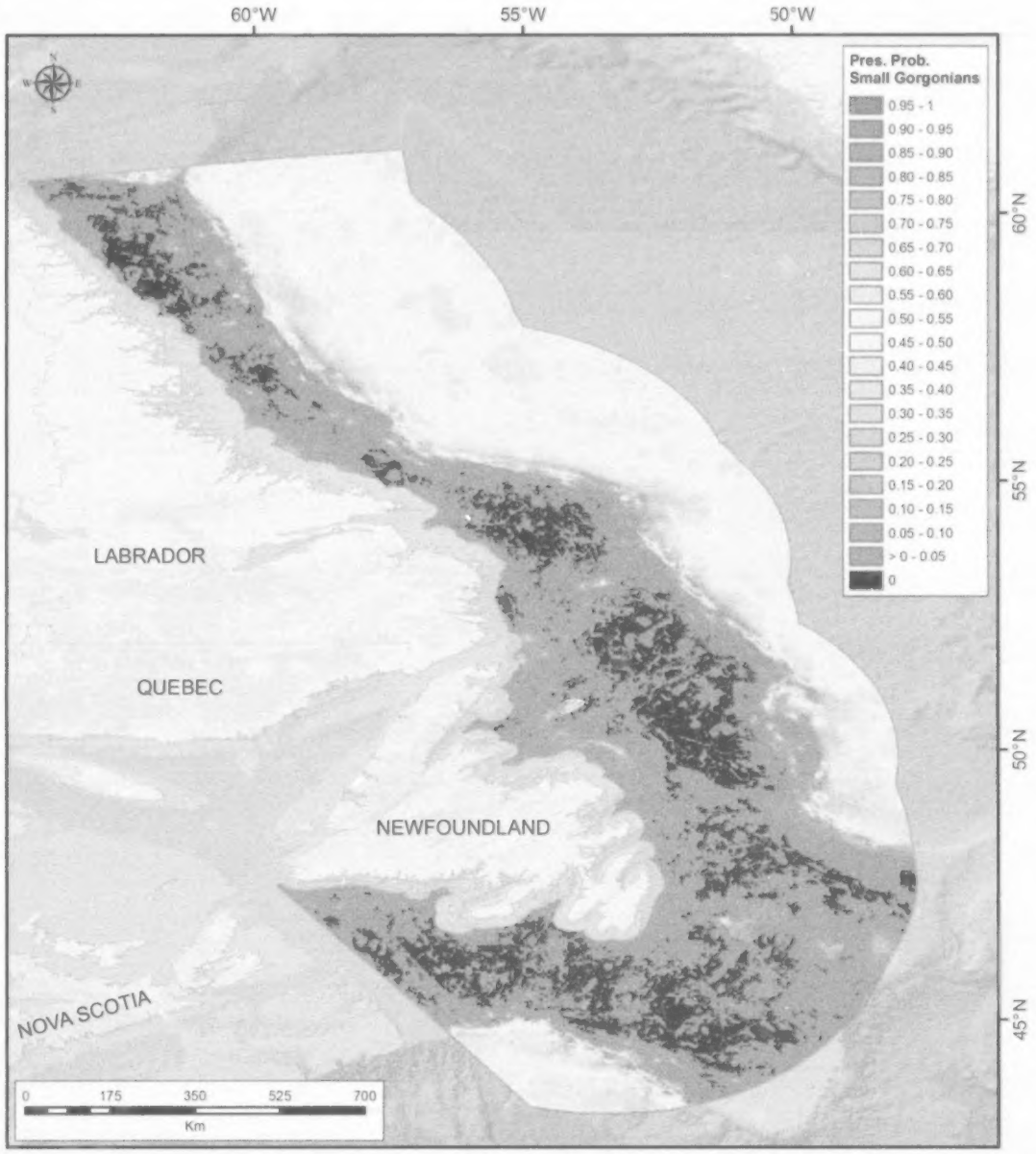

Figure 54. Predictions of presence probability from the unbalanced random forest model of small gorgonian corals presence and absence data collected from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

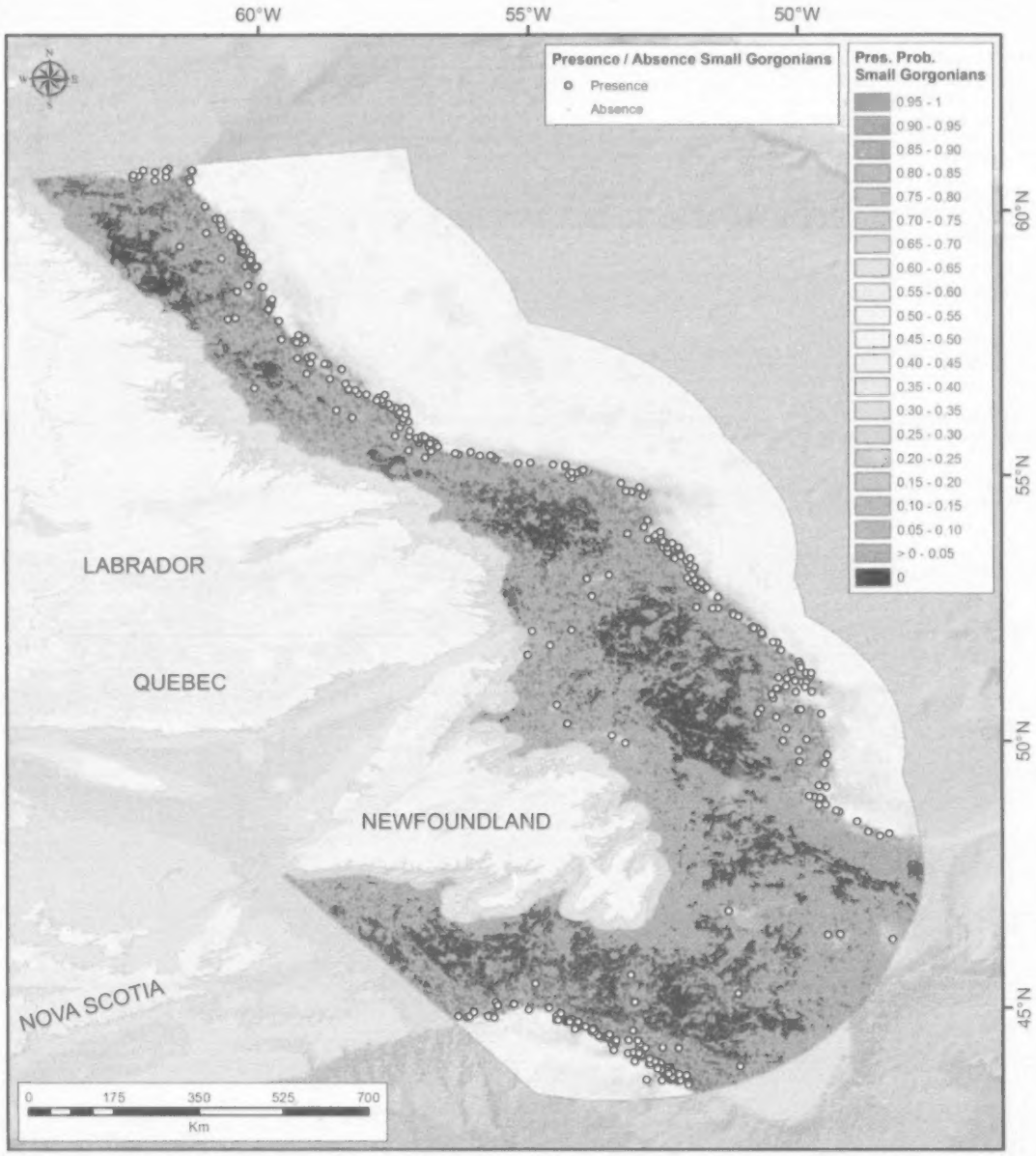

Figure 55. Presence and absence observations and predictions of presence probability from the unbalanced random forest model of small gorgonian coral presence and absence data collected from DFO multispecies surveys and DFO/industry northern shrimp surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

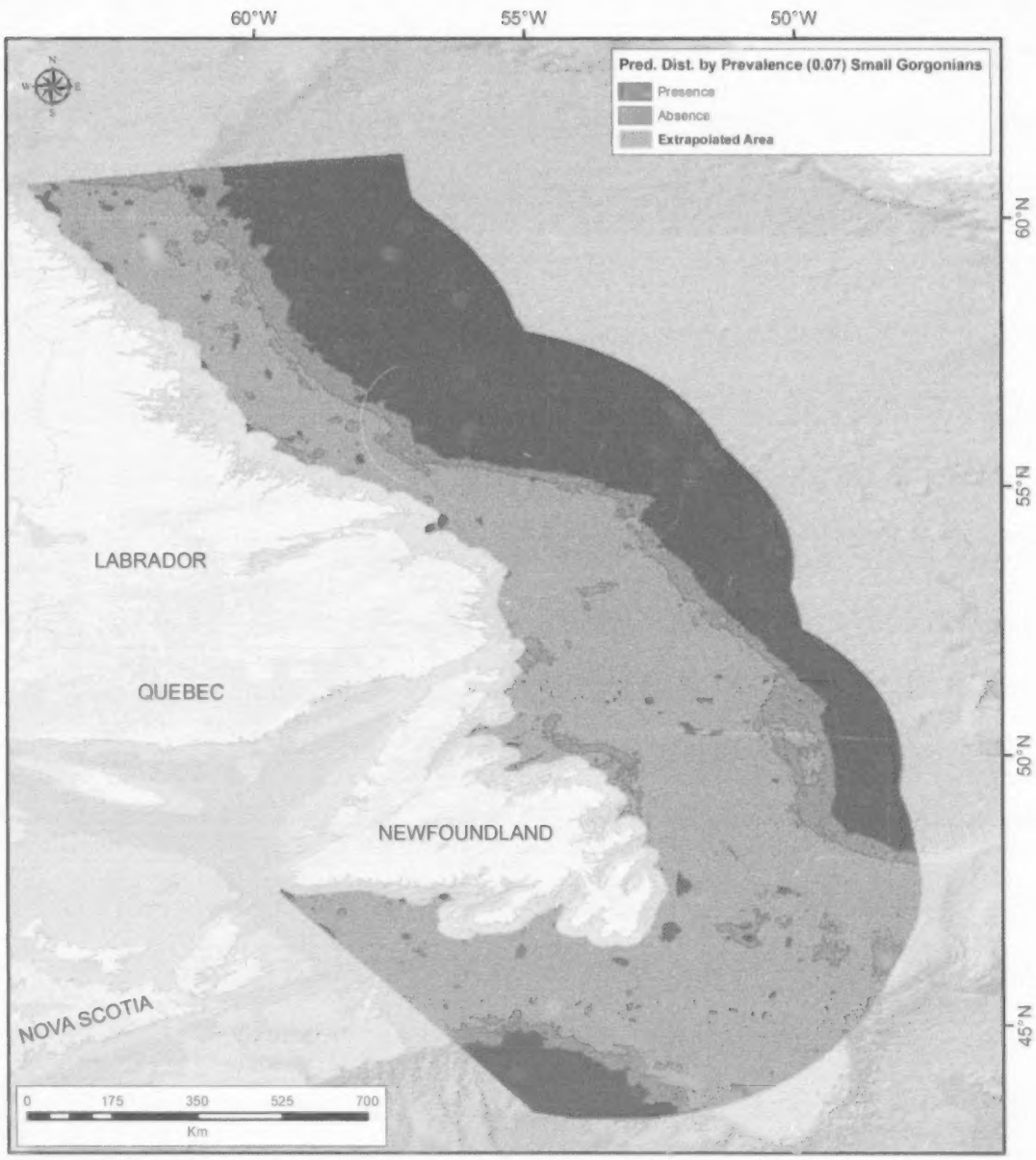

Figure 56. Predicted distribution (Pred. Dist.) of small gorgonian corals in the Newfoundland and Labrador Region based on the prevalence threshold of 0.07 of small gorgonian presence and absence data used in Model 2. Also shown are the areas of model extrapolation. Grey polygon may appear dark red or blue when overlain on the presence-absence surface.

Of the 66 environmental variables used in the model, Depth (non-interpolated variable) was the most important for the classification of the small gorgonian coral presence and absence data (Figure 57). This variable was followed importance by Bottom Salinity Average Minimum, Slope, and Bottom Salinity Average Range. Partial dependence of the small gorgonian coral

presence and absence data on the top 6 predictor variables is shown in Figure 58. The probability of small gorgonians rapidly increased at ~500 m depth and at Bottom Salinity Average Minimum value of ~34.5.

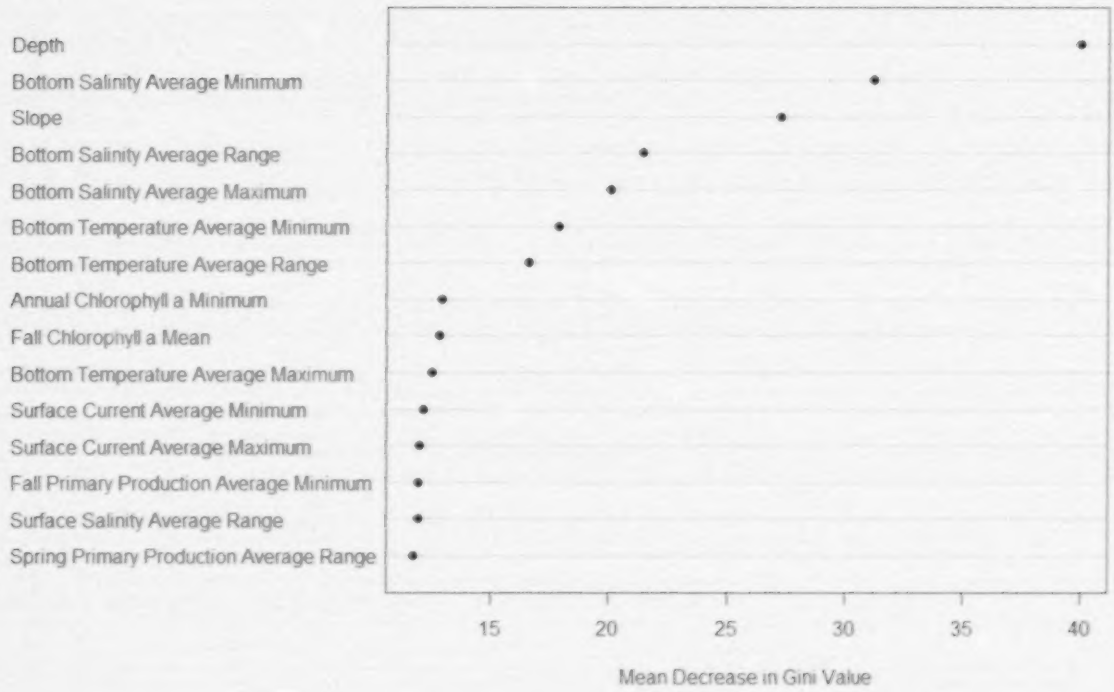

Figure 57. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini Value of the unbalanced random forest model of small gorgonian coral presence and absence data within the Newfoundland and Labrador Region. The higher the Mean Decrease in Gini Value the more important the variable is for predicting the response data.

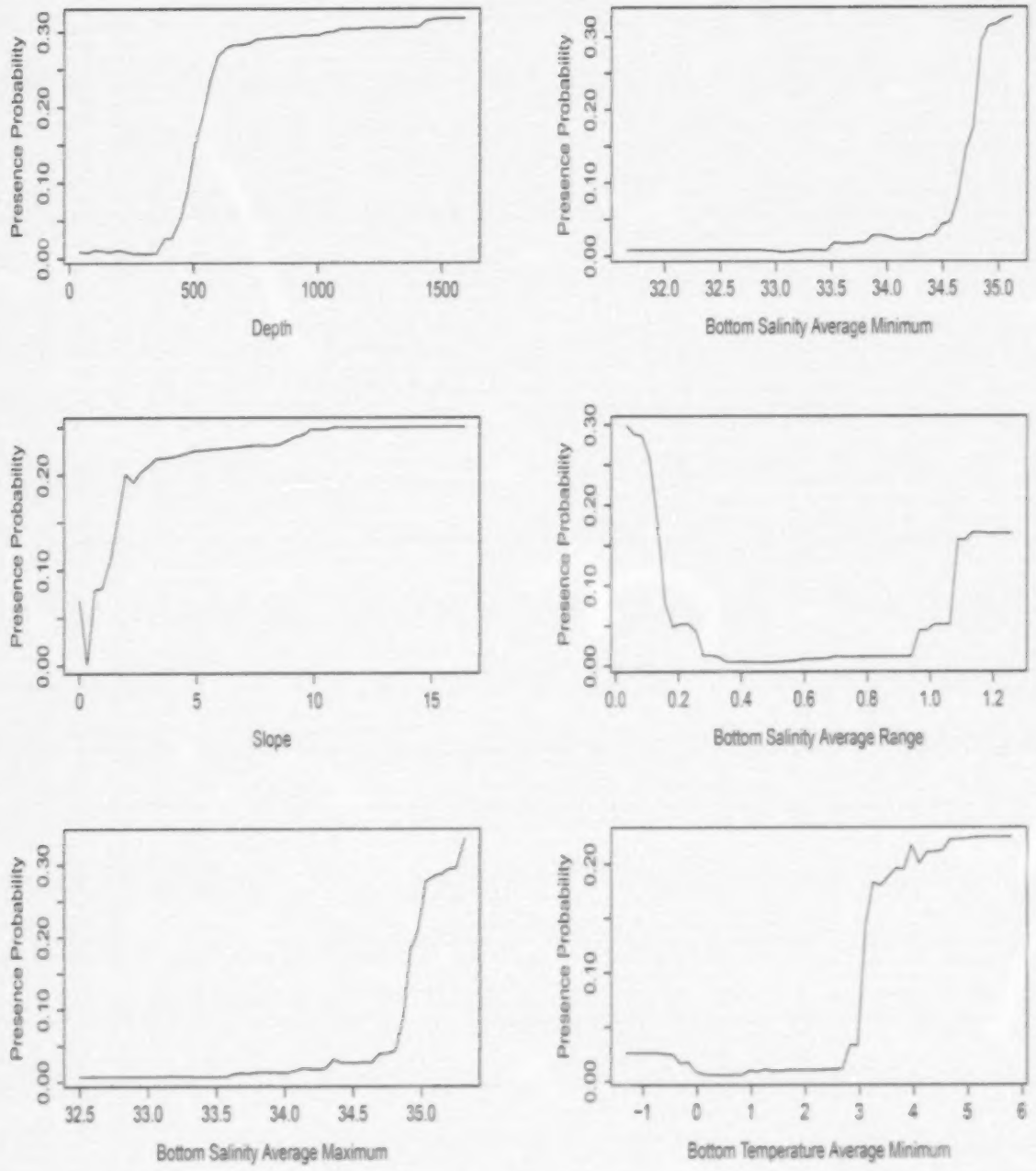

Figure 58. Partial dependence plots of the top 6 predictors from the unbalanced random forest model of small gorgonian coral presence and absence data collected within the Newfoundland and Labrador Region, ordered left to right from the top. Predicted presence probabilities are shown on the *y*-axis of each graph.

Model Selection

The random forest model using all available small gorgonian coral records and an unbalanced species prevalence and threshold equal to 0.07 (Model 2) was chosen as the best predictor of small gorgonian coral distribution in the Newfoundland and Labrador Region. Model 1 (balanced species prevalence) was considered a poor predictor of presence probability of small gorgonians due to its exaggeration of high presence probability beyond the location of presence data, particularly along the slopes. This phenomenon is likely due to random down-sampling of the absence data.

Prediction of Small Gorgonian Coral Biomass Using Random Forest

The accuracy measures of the regression random forest model on mean small gorgonian coral biomass per grid cell are presented in Table 18. The highest R^2 value was 0.267, while the average was 0.108 ± 0.080 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.021 ± 0.013 SD. The percent variance explained for most of the folds was negative, indicating that the model had no predictive power.

Figures 59 and 60 show the prediction surface of small gorgonian coral biomass per grid cell. The majority of the spatial extent was predicted to have low (> 0 – 0.01 kg) small gorgonian biomass. The highest predicted biomass (up to 1.48 kg) occurred on the slope southwest of Grand Bank.

Table 18. Accuracy measures from 10-fold cross validation of random forest model of average of small gorgonian coral biomass (kg) per grid cell recorded from DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.046	0.090	0.032	-1.51
2	0.050	0.086	0.031	-0.89
3	0.053	0.040	0.014	0.68
4	0.135	0.042	0.015	0.32
5	0.267	0.054	0.019	-7.07
6	0.219	0.038	0.014	-0.85
7	0.075	0.025	0.009	-0.84
8	0.132	0.024	0.009	-1.90
9	0.064	0.049	0.018	-2.45
10	0.039	0.136	0.049	2.20
Mean	**0.108**	**0.058**	**0.021**	**-1.23**
SD	**0.080**	**0.035**	**0.013**	**2.46**

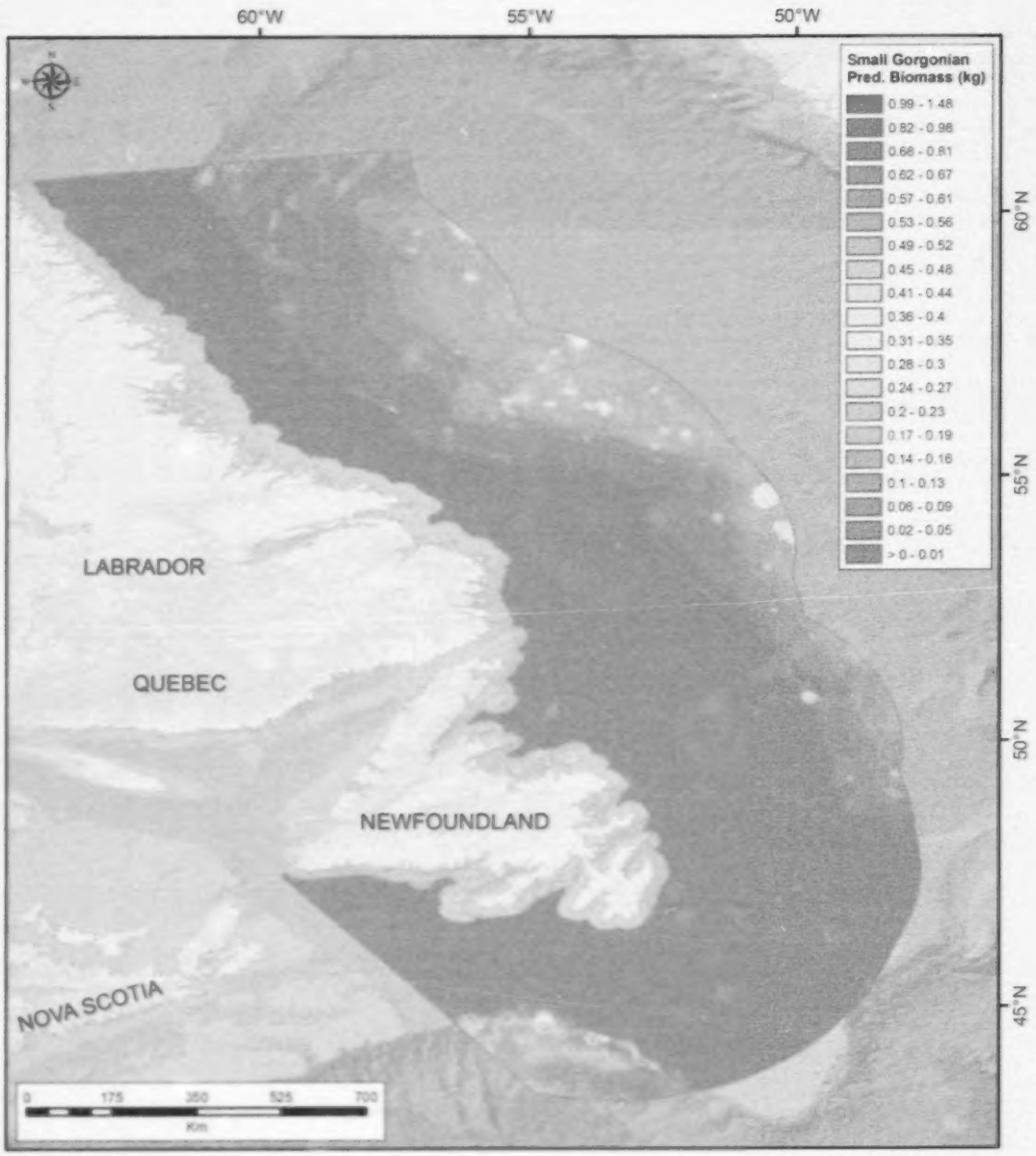

Figure 59. Predictions of biomass (kg) of small gorgonian corals from catch recorded in DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015.

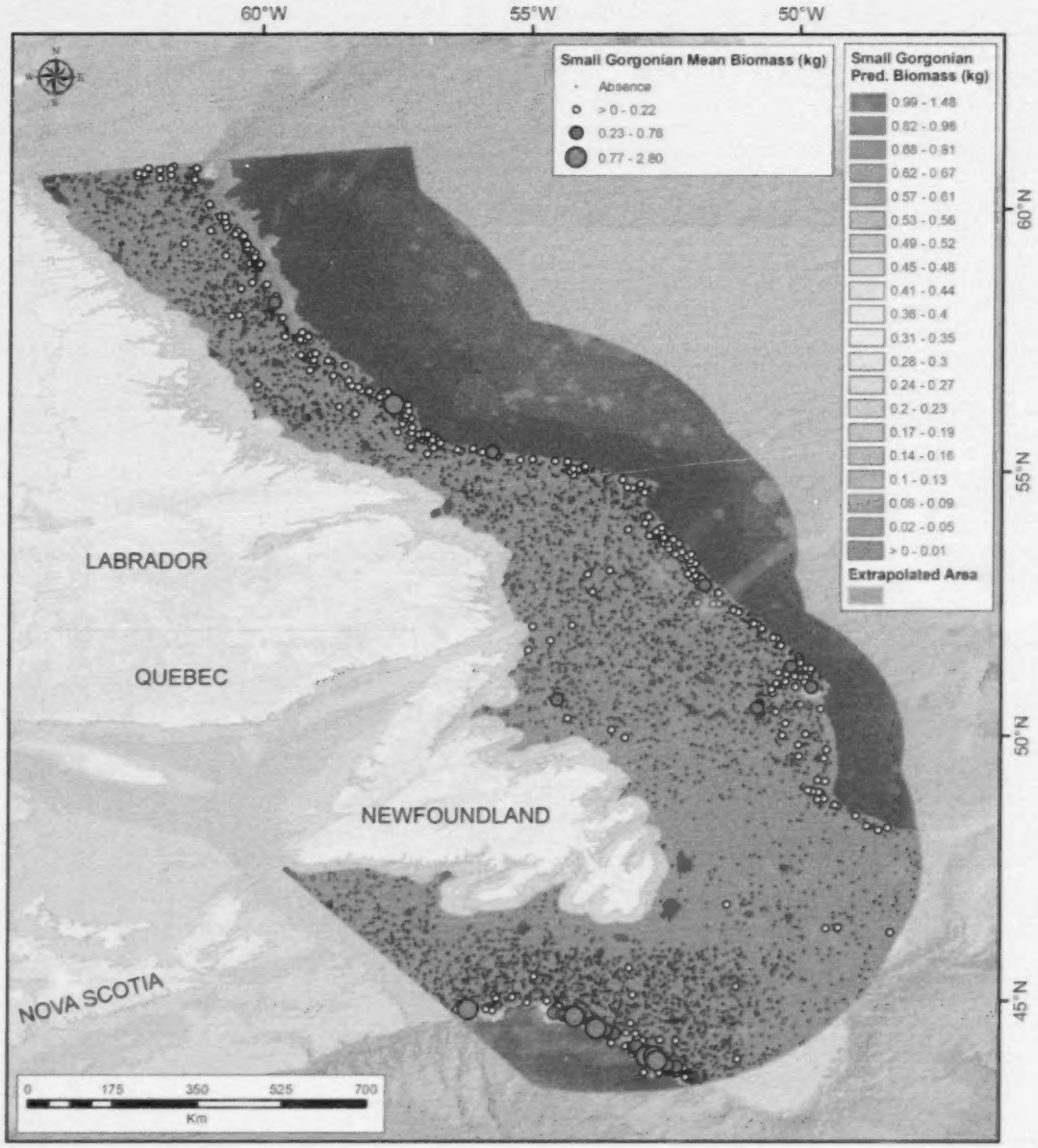

Figure 60. Predicted biomass (kg) of small gorgonian corals from catch recorded in DFO multispecies surveys, DFO/industry northern shrimp surveys, and Spanish groundfish trawl surveys conducted in the Newfoundland and Labrador Region between 2003 and 2015. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting small gorgonian coral biomass are shown in Figure 61. Fall Chlorophyll *a* Range was the most important variable in the model. Prior to spatial interpolation, this variable displayed a severely right-skewed distribution with outlying data in the upper range (Guijarro et al., in prep.). Examination of the Q-Q plot revealed

a weak spatial pattern in data points over- and under-predicted by a normal distribution, with under-predicted points taking precedence on the banks of Labrador. Fall Chlorophyll *a* Range was followed by Surface Temperature Average Minimum, Surface Temperature Average Range and Slope. The partial dependence of small gorgonian coral biomass on the top 6 most important variables is shown in Figure 62. Predicted biomass was the highest between 2 and 12 mg m^{-3} along the gradient in Fall Chlorophyll *a* Range. Values in this range were scattered across the study extent with no real spatial pattern. The fit between predicted and observed values in the kriging model was relatively poor, with under-predictions of values between 2 and 12 mg m^{-3}. Some points could therefore be predicted lower than their true values and slightly outside the range of highest predicted biomass as identified in the partial plot (Figure 62).

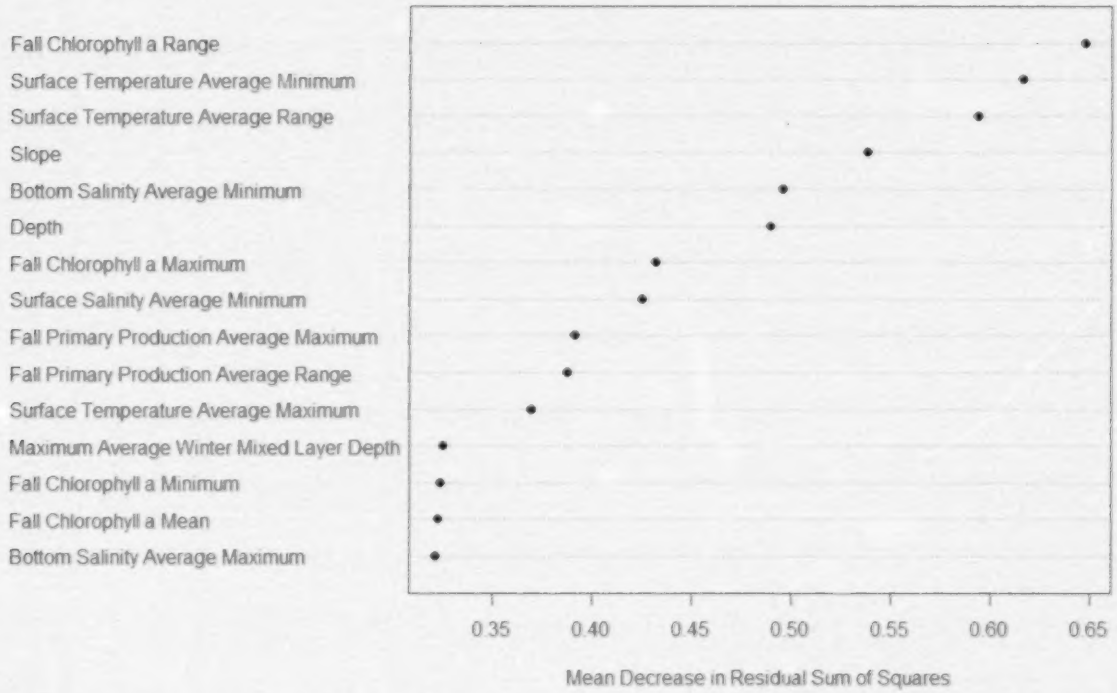

Figure 61. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on small gorgonian coral mean biomass data per grid cell. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

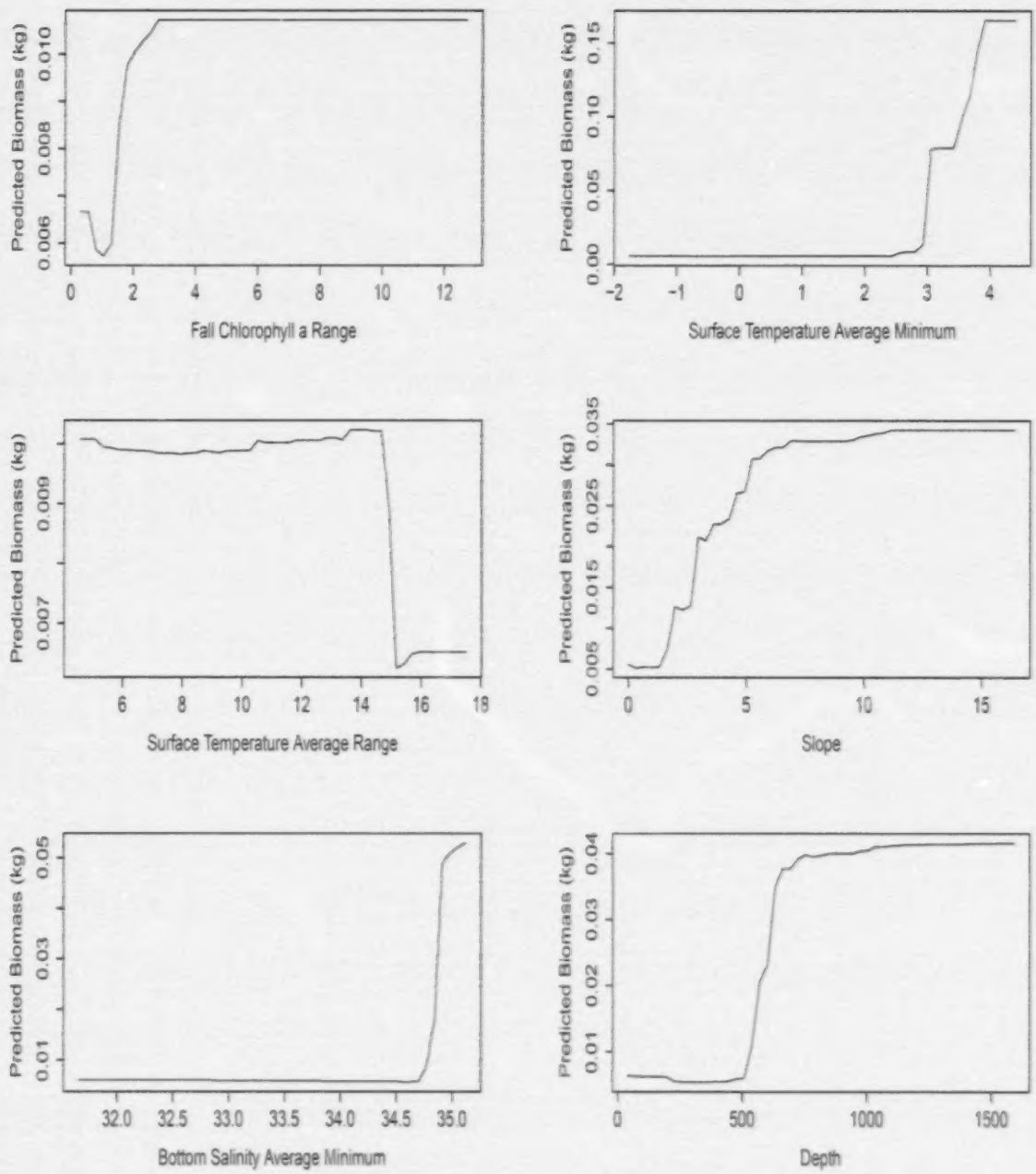

Figure 62. Partial dependence plots of the top six predictors from the random forest model of small gorgonian coral biomass collected within the Newfoundland and Labrador Region between 2003 and 2015, ordered left to right from the top. Predicted biomass is shown on the *y*-axis of each graph.

DISCUSSION

The species distribution models in this report were made possible through data collected over the years from numerous biogeographic/taxonomic surveys of corals and sponges in the Newfoundland and Labrador region. In this study, the sampling time series varied by taxonomic group, i.e., sponges (1995 - 2015) and corals (2003 - 2015). Standardized sampling and processing protocols employed on DFO multispecies survey and Observer Program vessels has produced georeferenced sample data throughout the region (Wareham et al., 2007, Wareham, 2009, Gilkinson and Edinger, 2009, *inter alia*). While taxonomic advances for coral taxa exceeds that for sponges, significant advancement is being made in deep-water sponge taxonomy, thus improving future predictive distribution modeling for this important benthic taxon.

This study is the first to use random forest modelling to predict the distribution of several sensitive deep-sea benthic invertebrate groups in the Newfoundland and Labrador Region. Table 19 shows a summary of the accuracy measures for the selected presence-absence and biomass random forest models of each taxonomic group. Presence-absence models for most groups performed well, with cross-validated AUC values ranging from 0.786 to 0.926. Similar performance of random forest models on the same taxonomic groups in the Maritimes Region was found by Beazley et al. (2016a), with models on sponges producing the lowest accuracy measures. This was attributed to the low taxonomic resolution of this group (phylum level) and the inclusion of both shelf and deep-water sponge species with preference for different habitats. In this report, performance of models on large gorgonians was comparable to that of sponges, suggesting that other factors may also be contributing.

Table 19. Summary of the mean accuracy measures for selected presence-absence models and biomass models for each of the four taxonomic groups. NRMSE = Normalized Root-Mean-Square-Error.

	Presence-absence			Biomass	
Taxon	**AUC**	**Sensitivity**	**Specificity**	**R^2**	**NRMSE**
Sponges (Porifera)	0.786	0.729	0.704	0.386	0.026
Sea Pens (Pennatulacea)	0.926	0.847	0.844	0.376	0.018
Large Gorgonian Corals	0.806	0.726	0.766	0.203	0.017
Small Gorgonian Corals	0.859	0.800	0.800	0.108	0.021

We have found that classification random forest models generated using all presence and absence data (i.e., unbalanced species prevalence) and a threshold equal to species prevalence produced the most realistic presence probability prediction surfaces These results are consistent with observations from RF models performed on sponge and coral groups in the DFO Maritimes Region, where there was highly unbalanced input data (see Beazley et al. 2016a). Random down-sampling of the absence data in such cases often resulted in gross extrapolation of high presence probability beyond the locations of presence observations. Beazley et al. (2016a) noted that this was particularly true in instances when the response data were imbalanced with a much greater number of absences than presences and down-sampling of the absence data was biased across the study area. In these instances, stratifying the down-sampling by spatial or sampling-effort strata may help reduce exaggerated predictions of presence probability and improve model

performance (but see Freeman et al., 2012). Nonetheless, our results may help guide future applications of random forest modelling by providing insight into which methods are appropriate based on the properties of the training data.

The random forest models worked well at interpolating predictions between data observations and extrapolating within the shallower (< 2000 m depth) portion of the study extent. However, the Newfoundland and Labrador Region extends out to the Canadian EEZ to approximately 4360 m depth, far outside the depth distribution of the training data. These deeper areas are also likely to have different physical environmental conditions, such as temperature and salinity, from those used to train the model. Random forest models give a constant value for inputs falling under each tree leaf, but when extrapolating outside of their training domain, they use the predicted outcome from the nearest point at which there is training data (Breiman et al., 1984). For true extrapolation, the random forest algorithm would need to learn the functional relationship between the response and environmental conditions at those locations. Therefore, we are not confident of the model extrapolations to depths beyond the limit of the training data. Sponges, sea pens, and gorgonian corals can be found at such depths and so the model may be helpful in guiding research surveys to perform such validation. Validation of models within the shallow portion of the study extent (see Appendix 2) showed good spatial congruence between presence probability and the distribution of observer records for most taxonomic groups. However, som older observer records (1985 to 2001) of sponges on Grand Bank were predicted as absence by the model (see Figure A2.2).

Table 20 summarizes the top predictor variable for each of the coral and sponge taxa for the random forest Model 2 (unbalanced) presence-absence and biomass models. For the presence-absence models depth was a key predictor variable for all taxa, including the sponges where it was the second highest ranking predictor after Fall Primary Production Average Maximum. In contrast, biomass was best predicted by variables related to food supply, which is consistent with the results for the Gulf of St. Lawrence (Murillo et al., 2016a).

Table 20. Summary of the top predictor variables for the best fit presence-absence models and biomass models for each of the four taxonomic groups.

Taxon	Top Predictor Pres-Abs	Top Predictor Biomass
Sponges (Porifera)	Fall Primary Production Average Maximum	Summer Primary Production Average Minimum
Sea Pens (Pennatulacea)	Depth	Maximum Average Winter Mixed Layer Depth
Large Gorgonian Corals	Depth	Summer Primary Production Average Minimum
Small Gorgonian Corals	Depth	Fall Chlorophyll *a* Range

The random forest and generalized additive models (GAMs) predicted similar areas of high biomass of the coral and sponge groups (see Appendix 1). For some groups (e.g. small gorgonian corals, see Figures A1.8 and A1.9), GAMs provided better predictions of biomass along the

slopes of Newfoundland and Labrador compared to those of random forest. However, the GAMs did not serve to resolve predictions in the deeper waters beyond the slope that are considered extrapolated by the random forest models. An exception was the sea pen GAM model, which predicted a localized area of high biomass on the slope southwest of Grand Bank, which was predicted to have only low to medium biomass over a broader area by the random forest model.

Knudby et al. (2013a) used random forest to predict the distribution of several sponge species and sponge grounds in the northwest Atlantic including the DFO Newfoundland and Labrador Region. This model predicted sponge grounds to occur with moderate probability along the slopes of Newfoundland and Labrador (Figure 63), with little to no presence probability on the continental shelf. These models were run on catches above a biomass threshold (200 kg) which served to distinguish habitat dominated by large structure-forming *Geodia* sponges which congregate along the continental slopes, from smaller sponge species that dominate the shelf (Knudby et al., 2013a). In order to directly compare our results with Knudby et al. (2013a) we first had to rerun our sponge (Porifera) Model 2 using only trawl catchers greater or equal to 200 kg over a similar spatial extent to that of the NL subarea. The two models performed similarly (AUC= 0.946 in Knudby et al., 2013a; AUC= 0.991 herein) despite our model using primary production variables that were not available to Knudby et al. (2013a) and that ranked high in the models, and having additional response data records. However the results were not identical and our model showed higher probability of sponge grounds predicted to occur along the slopes off Labrador and the Northeast Newfoundland Shelf. Both models showed little to no occurrence on the continental

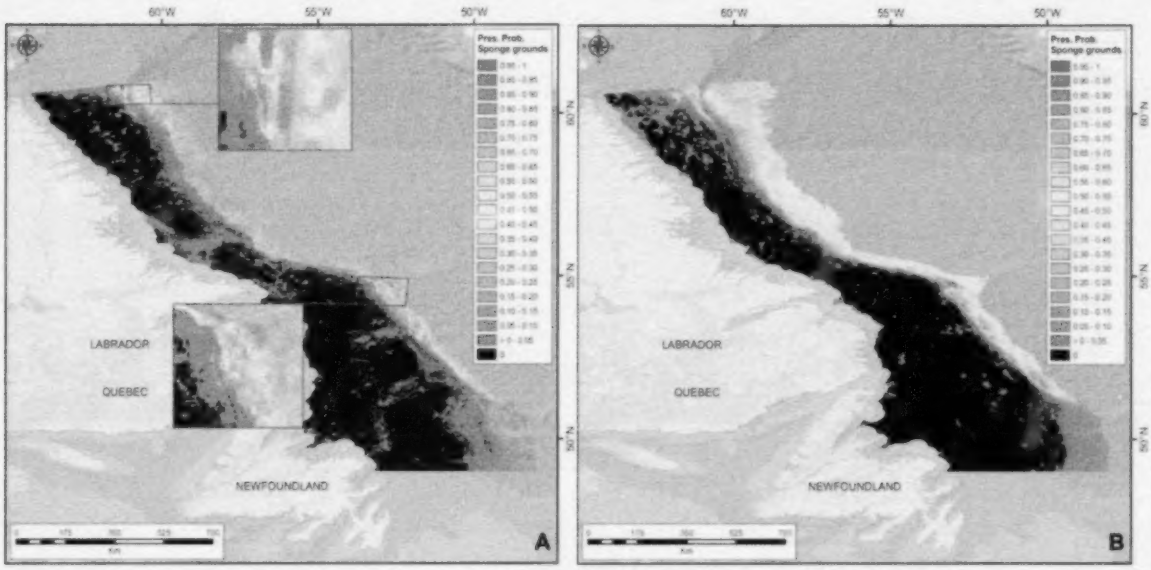

Figure 63. Spatial concordance between A) the predicted presence probability of sponge grounds from the NL subarea of Knudby et al. (2013a) clipped to the 20-km land buffer used in our study, and B) our model results using an unbalanced presence-absence RF model fit to sponge catches greater than 200 kg over the same spatial extent.

shelf (Figure 63). The area of highest sponge ground presence probability from Knudby et al. (2013a) was located on the slope off Saglek Bank in northern Labrador and on the slope off Labrador/Northeast Newfoundland Shelf. Our model also predicted high probability of occurrence of sponges in those areas.

The validity of any SDM is partially driven by choice/availability of predictor environmental variables. In this study, a total of 66 environmental predictor variables were considered based on suspected/known importance for explaining the distribution (i.e. presence-absence) of benthos including deep-water corals and sponges. It is noted that the order of importance of predictor variables differed between sponge/coral groups including presence-absence and biomass data within a given group. This may highlight the complexity of interacting environmental factors controlling the distribution, in space and time, of these taxa. It is also noted that our understanding of biogeographic patterns within each of these larger species groups will only improve with increased sampling effort (preferentially non-destructive) and knowledge of how the species respond to the key environmental variables. Edinger et al. (2011) recommended in a qualitative study, that predictive models of coral distribution should consider Quaternary and surficial geology. While currently this information is not available throughout the region there has been expansion of multibeam coverage in the NL region, and in future, such habitat variables could be included in predictive models. Sediment grain size in the NAFO Regulatory Area has been used to produce interpolated and interpreted surficial geological maps, however the interpolation surfaces based on a large number of samples did not elucidate the fine-scale patchiness that can be found on the seafloor and the results were only suitable for post-hoc characterization (Murillo et al., 2016b). Interpolations of such data performed well on the Sackville Spur (see Beazley et al., 2015), a relatively small-scale area compared to the spatial extent used in this report, with relatively homogenous surficial characteristics. Additional information such as bathymetry, slope, and multibeam backscatter used as covariates in spatial interpolation models may help improve interpolation of grain size data (Leecaster, 2003; Jerosch, 2013). Nonetheless, surficial characteristics at the resolution of the SDM maps (1 km^2) can be heterogenous (cf. Cuff et al., 2015; Rincón and Kenchington, 2016) and although fine-scale seafloor mapping could be useful to resolve fine-scale models (Beazley et al., 2013), at regional scales applicable to management issues, such information is not likely to be informative. We note that variables such as shear, slope, depth and bottom currents will be correlated with many aspects of surficial geology (Li et al., 2016), which is why the models perform well despite not having direct information on substrate type.

SDMs identify potential species distribution and can indicate areas for future restoration initiatives towards the implementation of the Policy for Managing the Impact of Fishing on Sensitive Benthic Areas. This policy was developed by DFO in 2009 to ensure Canadian fisheries are conducted in a manner that supports marine conservation and sustainable resource use within and outside Canada's 200 nautical mile EEZ. These models provide continuous surfaces of presence and biomass that can fill in gaps in survey coverage and extrapolate to a certain degree to areas outside of the surveys. Combined with kernel density analysis (Kenchington et al., 2016) SDM can be used to refine significant benthic area polygons produced by the former by clipping boundaries to more probabilistic borders.

ACKNOWLEDGMENTS

This project was funded in part by a one year project under DFO's Strategic Program for Ecosystem-Based Research and Advice (SPERA) to EK and through financial support by DFO's Oceans Management Program, Newfoundland and Labrador Region. We thank Annette Power for the latter contribution and guidance in the preparation of this report. We thank C. Rooper (NMFS - RACE Division, Washington, USA) and K. Tanaka (U of Maine, Maine, USA) for their constructive comments on SDM and GAMs.

REFERENCES

Baker, K.D., Haedrich, R.L., Snelgrove P.V.R., Wareham, V.E., Edinger, E.N., and Gilkinson, K.D. 2012a. Small-scale patterns of deep-sea fish distributions and assemblages of the Grand Banks, Newfoundland continental slope. Deep-Sea Res. I 65: 171–188.

Baker, K.D., Wareham, V.E., Snelgrove, P.V.R., Haedrich, R.L., Fifield, D.A., Edinger, E.N., and Gilkinson, K.D. 2012b. Distributional patterns of deep-sea coral assemblages in three submarine canyons off Newfoundland, Canada. Mar. Ecol. Prog. Ser. 445: 235–249.

Barrio-Froján, C.R.S., MacIsaac, K.G., McMillan, A.K., Del Mar Sacau Cuadrado, M., Large, P., Kenny, A.J., Kenchington, E. and De Cárdenas González, E. 2012. An evaluation of benthic community structure in and around the Sackville Spur closed area (Northwest Atlantic) in relation to the protection of vulnerable marine ecosystems. ICES J. Mar. Sci. 69: 213–222.

Beazley, L.I., and Kenchington, E.L. 2015. Epibenthic Megafauna of the Flemish Pass and Sackville Spur (Northwest Atlantic) Identified from *In Situ* Benthic Image Transects. Can. Tech. Rep. Fish. Aquat. Sci. 3127: v + 496p.

Beazley, L.I., Kenchington, E., Murillo, F.J., and Sacau, M. 2013. Deep-sea sponge grounds enhance diversity and abundance of epibenthic megafauna in the Northwest Atlantic. ICES J. Mar. Sci. 70: 1471–1490.

Beazley, L., Kenchington, E., Yashayaev, I., and Murillo, F.J. 2015. Drivers of epibenthic megafaunal composition in the sponge grounds of the Sackville Spur, northwest Atlantic. Deep-Sea Res. I 98: 102–114.

Beazley, L., Kenchington, E., Murillo, J., Lirette, C., Guijarro, J., McMillan, A., and Knudby, A. 2016a. Species Distribution Modelling of Corals and Sponges in the Maritimes Region for Use in the Identification of Sensitive Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3172: vi + 189p.

Beazley, L., Lirette, C., Sabaniel, J., Wang, Z., Knudby, A., and Kenchington, E. 2016b. Characteristics of Environmental Data Layers for Use in Species Distribution Modelling in the Gulf of St. Lawrence. Can. Tech. Rep. Fish. Aquat. Sci. 3154: viii + 357p.

Breiman, L. 2001. Random Forests. Mach. Learn. 45: 5–32.

Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J. 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, California.

Campbell, J.S., and Simms, J.M. 2009. Status Report on Coral and Sponge Conservation in Canada. Fisheries and Oceans Canada: vii + 87 p.

Chen, C., Liaw, A., and Breiman, L. 2004. Using Random Forest to learn imbalanced data. University of California, Berkeley. 12 p.

Cuff, A., Anderson, J.T., and Devillers, R. Comparing surficial sediments maps interpreted by experts with dual-frequency acoustic backscatter on the Scotian shelf, Canada. Cont. Shelf Res. 110: 149–161.

Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., and Lawler, J.J. Random Forest for classification in ecology. Ecology 88: 2783–2792.

DFO. 2013. Identification of Additional Ecologically and Biologically Significant Areas (EBSAs) within the Newfoundland and Labrador Shelves Bioregion. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2013/048.

Dunn, P. K., and Smyth, G. K. 1996. Randomized quantile residuals. Journal of Computational and Graphical Statistics, 5: 236-244.

Edinger, E.N., and Sherwood, O.A. 2012. Applied taphonomy of gorgonian and antipatharian corals in Atlantic Canada: experimental decay rates, field observations, and implications for assessing fisheries damage to deep-sea coral habitats. N. Jb. Geol. Paläont. Abh. 265/2, 199–218.

Edinger, E.N., Wareham, V.E., and Haedrich, R.L. 2007. Patterns of groundfish diversity and abundance in relation to deep-sea coral distributions in Newfoundland and Labrador waters. Bull. Mar. Sci. 81, Suppl. 1: 101–122.

Edinger, E.N., Sherwood, O.A., Piper, D.J.W., Wareham, V.E., Baker, K.D., Gilkinson, K.D., and Scott, D.B. 2011. Geological features supporting deep-sea coral habitat in Atlantic Canada. Cont. Shelf Res 31: S69–S84.

Elith, J., Kearney, M., and Phillips, S. 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1: 330–342.

ESRI. 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Evans J.S., Murphy, M.A., Holden, Z.A., and Cushman, S.A. 2011. Modeling Species Distribution and change Using Random Forests. *In* Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications. Edited by C.A Drew, Y.F. Wiersma, and F. Huettmann. Springer, New York. pp. 139-159.

FAO. 2009. International Guidelines for the Management of Deep-sea Fisheries in the High Seas. FAO, Rome. 73 p.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recog. Lett. 27: 861–874.

Freeman, E.A., Moisen, G.G., and Frescino, T.S. 2012. Evaluating effectiveness of down-sampling for stratified designs and unbalanced prevalence in Random Forest models of tree species distributions in Nevada. Ecol. Model. 233: 1–10.

Gilkinson, K., and Edinger, E. (Eds.) 2009. The ecology of deep-sea corals of Newfoundland and Labrador waters: biogeography, life history, biogeochemistry, and relation to fishes. Can. Tech. Rep. Fish. Aquat. Sci. 2830: vi + 136 p.

Guijarro, J., Beazley, L., Lirette, C., Wang, Z., and Kenchington, E. In Prep. Characteristics of Environmental Data Layers for Use in Species Distribution Modelling in the Newfoundland and Labrador Region. Can. Tech. Rep. Fish. Aquat. Sci.

Guisan, A., and Zimmerman, N.E. 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135: 147–186.

Hamel, J.F., Sun, Z., and Mercier, A. 2010. Influence of size and seasonal factors on the growth of the deep-sea coral *Flabellum alabastrum* in mesocosm. Coral Reefs 29: 521–525.

Hamoutene, D., Puestow, T., Miller-Banoub, J., and Wareham, V. 2007. Main lipid classes in some species of deep-sea corals in the Newfoundland and Labrador region (Northwest Atlantic Ocean). Coral Reefs DOI 10.1007/s00338-007-0318-7.

Hanberry, B.B. and He, H.S. 2013. Prevalence, statistical thresholds, and accuracy assessment for species distribution models. Web Ecol. 13: 13–19.

Hastie, T., and Tibshirani, R. 1986. Generalized additive models. Stat. Sci. 1: 297-318.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. 2005. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition. Springer+Verlag.

Herrick, K.K., Huettmann, F., and Lindgren, M.A. 2013. A global model of avian influenza prediction in wild birds: the importance of northern region. Vet. Res. 44:42.

Jerosch, K. Geostatistical mapping and spatial variability of surficial sediment types on the Beaufort Shelf based on grain size data. J. Mar. Syst. 127: 5–13.

Jiménez-Valverde, A., and Lobo, J. M. 2006. The ghost of unbalanced species distribution data in geographical model predictions. Divers. Distrib. 12: 521–524.

Kenchington, E. 2014. A General Overview of Benthic Ecological or Biological Significant Areas (EBSAs) in Maritimes Region. Can. Tech. Rep. Fish. Aquat. Sci. 3072: iv + 45 p.

Kenchington, E., Murillo, F.J., Lirette, C., Sacau, M., Koen-Alonso, M., Kenny, A., Ollerhead, N., Wareham, V., and Beazley, L. 2014. Kernel density surface modelling as a means to identify significant concentrations of vulnerable marine ecosystem indicators. PLoS ONE 10(1): e0117752. doi:10.1371/journal.pone.0117752

Kenchington, E., Lirette, C., Murillo, F.J., Beazley, L., Guijarro, J., Wareham, V., Gilkinson, K., Koen Alonso, M., Benoît, H., Bourdages, H., Sainte-Marie, B., Treble, M., and Siferd, T. 2016. Kernel Density Analyses of Coral and Sponge Catches from Research Vessel Survey Data for Use in Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3167: viii + 207p.

Knudby, A., Kenchington, E., and Murillo, F.J. 2013a. Modelling the distribution of *Geodia* sponges and sponge grounds in the Northwest Atlantic. PLoS One 8, e82306. http://dx.doi.org/10.1371/journal.pone.0082306.

Knudby, A., Lirette, C., Kenchington, E., and Murillo, F.J. 2013b. Species distribution models of black corals, large gorgonian corals, and sea pens in the NAFO Regulatory Area. NAFO SCR Doc 13/78, Ser. No N6276. 17 p.

Kuhn, M. and Johnson, K. 2013. Applied Predictive Modeling. New York: Springer Science + Business Media.

Leecaster, M. 2003. Spatial analysis of grain size in Santa Monica Bay. Mar. Environ. Res. 56: 67–78.

Li, J., Tran, M., and Siwabessy, J. 2016. Selecting optimal Random Forest predictive models: a case study on predicting the spatial distribution of seabed hardness. PloS One 11(2): e0149089. doi:10.1371/journal.pone.0149089.

Liaw, A., and Wiener, M. 2002. Classification and regression by randomForest. R News, 2: 18–22.

Liu, C., Berry, P.M., Dawson, T.P., and Pearson, R.G. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28: 385–393.

McPherson, J.M., Jetz, W., and Rogers, D.J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artifact? J. Appl. Ecol. 41: 811–823.

Matthiopoulos, J. 2011. How to be a Quantitative Ecologist: The 'A to R' of Green Mathematics and Statistics. Wiley, Chichester, West Sussex.

Mercier, A., Sun, Z., and Hamel, J.F. 2011a. Reproductive periodicity, spawning and development of the deep-sea scleractinian coral *Flabellum angulare*. Mar. Biol. 158: 371–380.

Mercier, A., Sun, Z., Baillon, S.B., and Hamel, J.F. 2011b. Lunar rhythms in the deep sea: evidence from the reproductive periodicity of several marine invertebrates. J. Biol. Rhythms. 26(1): 82–6. doi: 10.1177/0748730410391948.

Miller, D.L., Rexstad, E., Burt, L., Bravington, M.V., and Hedley, S. 2015. Package 'dsm'. 26 p.

Murillo, F.J., Durán Muñoz, P., Altuna, A., and Serrano, A. 2011. Distribution of deep-water corals of the Flemish Cap, Flemish Pass, and the Grand Banks of Newfoundland (Northwest Atlantic Ocean): interaction with fishing activities. ICES J. Mar. Sci. 68: 319–332.

Murillo, F.J., Durán Muñoz, P., Cristobo, F.J., Ríos, P., González, C., Kenchington, E., and Serrano, A. 2012. Deep-sea Sponge Grounds of the Flemish Cap, Flemish Pass and the Grand Banks of Newfoundland (Northwest Atlantic Ocean): distribution and species composition. Mar. Biol. Res. 8: 842–854.

Murillo, F.J., E. Kenchington, L. Beazley, C. Lirette, A. Knudby, J. Guijarro, H. Benoît, H. Bourdage, and B. Sainte-Marie. 2016a. Distribution Modelling of Sea Pens, Sponges, Stalked Tunicates and Soft Corals from Research Vessel Survey Data in the Gulf of St. Lawrence for Use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3170: vi + 132p.

Murillo, F.J., Serrano, A., Kenchington E., and Mora, J. 2016b. Epibenthic assemblages of the Tail of the Grand Bank and Flemish Cap (northwest Atlantic) in relation to environmental parameters and trawling intensity. Deep Sea Res. I 109: 99–122.

NAFO. 2013. Report of the 6th meeting of the NAFO Scientific Council Working Group on Ecosystem Science and Assessment (WGESA). NAFO SCS, Doc. 13/24, Serial No. N6277. 208 p.

NAFO. 2015. Conservation and enforcement measures. NAFO/FC, Doc. 15/01, Serial No. N6409. 134 p.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rincón, B, and Kenchington, E. 2016. Spatial and temporal variation of benthic macrofauna on the eastern Scotian Shelf: association with juvenile Haddock (*Melanogrammus aeglefinus*) spatial structure and environmental drivers. Under Review.

Sherwood, O.A., and Edinger, E.N. 2009. Ages and growth rates of some deep-sea gorgonian and antipatharian corals of Newfoundland and Labrador. Can. J. Fish. Aquat. Sci. 66: 142–15.

Sherwood, O.A., Jamieson, R.E., Edinger, E.N., and Wareham, V.E. 2008. Stable C and N isotopic composition of cold-water corals from the Newfoundland and Labrador continental slope: Examination of tropic, depth and spatial effects. Deep-Sea Res. I 55: 1392–1402.

Sherwood, O., Lehmann, M., Schubert, C., Scott, D. and McCarthy, M. 2011. Nutrient regime shift in the western North Atlantic indicated by compound-specific δ15N of deep-sea gorgonian corals, Proc. Natl. Acad. Sci. U. S. A., 108, 1011–1015.

Shono, H. 2008. Application of the Tweedie distribution of zero-catch data in CPUE analysis. Fish. Res. 93: 154–162.

Sun, Z., Hamel, J.F. and Mercier, A. 2010. Planulation periodicity, settlement preferences and growth of two deep-sea octocorals from the northwest Atlantic. Mar. Ecol. Prog. Ser. 410: 71–87.

Templeman, ND. 2007. Placentia Bay-Grand Banks Large Ocean Management Area Ecologically and Biologically Significant Areas. Can. Sci. Advis. Sec. Res. Doc. 2007/052: iii + 15 p.

Tweedie, M.C.K. 1984. An index which distinguishes between some important exponential families. In: Ghosh, J.K., Roy, J. (Eds.), Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Indian Statistical Institute, Calcutta, pp. 579–604.

Wareham, V. E. 2009. Update on deep-sea coral distributions in the Newfoundland Labrador and Arctic regions, Northwest Atlantic. *In* The ecology of deep-sea corals of Newfoundland and Labrador waters: biogeography, life history, biogeochemistry, and relation to fishes. Edited by K. Gilkinson, and E. Edinger pp. 4-22. Can. Tech. Rep. Fish. Aquat. Sci. 2830: vi + 136p.

Wareham, V.E., and Edinger, E.N. 2007. Distributions of deep-sea corals in the Newfoundland and Labrador region, Northwest Atlantic Ocean. Bull. Mar. Sci. 81: 289-312.

Wareham, V.E., Ollerhead, N.E., and Gilkinson, K.D. 2010. Spatial Analysis of Coral and Sponge Densities with Associated Fishing Effort in Proximity to Hatton Basin (NAFO Divisions 2G-0B). DFO Can. Sci. Advis. Sec. Res. Doc. 2010/058, 46 pp.

Wood, S.N. 2006. Generalized additive models: an introduction with R. Chapman & Hall/CRC Press, Boca Raton, FL.

APPENDIX 1

Alternative Prediction Models- Generalized Additive Models for Predicting Coral and Sponge Biomass in the Newfoundland and Labrador Region

Given the fair to poor prediction of biomass by the random forest models, particularly in deep water, generalized additive models (GAMs; Hastie and Tibshirani, 1986) were developed to compare to the random forest results and to determine whether predictions could be improved for the areas considered as extrapolated by random forest models. A generalized additive model (Hastie and Tibshirani, 1986; 1990) is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates. GAM models follow this general structure:

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots$$

where $\mu_i \equiv E(Y_i)$ and $Y_i \sim$ some exponential family distribution. Y_i is a response variable, X_i^* is a row of the model matrix for any strictly parametric model components, θ is the corresponding parameter vector, and the f_j are smooth functions of the covariates, x_k (Wood, 2006). The model allows for somewhat flexible specification of the dependence of the response on the covariates. This flexibility provides potential for a better fit to the data than purely parametric models.

Two different approaches were used to select the predictor variables. In the first approach, highly correlated variables were identified and eliminated in order to increase interpretability of the models and to reduce the effects of collinear variables. This was done following the variable elimination procedure outlined in Knudby et al. (2013a). The Spearman's rank correlation coefficients between all predictor variables in the study area were calculated from all raster cells in the study area, and the two predictors with the highest correlation were then considered and one of them eliminated. This process was repeated until there were no variables remaining that were correlated higher than 0.7. Models generated using the variables selected with this approach are termed 'GAM 0.7 Variables' herein. The second approach involved selecting the top predictor variables identified in the random forest biomass models. This was done independently for each taxonomic group. Those variables with a higher influence in the RF models were identified by examining the importance plots and identifying those variables that fell above a natural break in the Mean Decrease in Residual Sum of Squares. Models generated with variables selected using this approach are termed 'GAM RF Variables' herein.

The Tweedie distribution (Tweedie, 1984) was utilized for each model. The Tweedie model is an expansion of a compound Poisson model derived from the stochastic process where the weight of the counted objects has a gamma distribution. This model has the advantage of handling the zero-catch data in a unified way and has shown to outperform the two-stage Delta lognormal model (Shono, 2008). The 'mgcv' package in R (Wood, 2006) was used to construct the GAMs.

Shrinkage smoothers were applied to each covariate in the form of a penalized cubic regression spline (s(variable, bs='cs')). Shrinkage smoothers allow the 'wiggliness' of each covariate to go to zero as required by the data (Matthiopoulos, 2011). Shrinkage smoothers are useful for variable selection, as such covariates remain in the model but have no effect on model predictions. For each model, autocorrelation in the residuals was determined by examining ACF

plots. When autocorrelation appeared substantial, latitude and longitude were included in the model as a tensor product (i.e. te(lat, long)).

Model performance was evaluated by assessing the goodness-of-fit statistic R^2, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), and the percent (%) deviance explained. Accuracy measures and model performance were compared between models generated using each set of predictor variables.

Residual plots to evaluate the fitness of the model were generated using the 'gam.check' function of the 'mgcv' package. However, an artifact of the link function shows exact zeros as a band along the residuals vs. linear predictor plot, making it difficult to see whether residuals show heteroskedasticity. In order to avoid this issue, randomized quantile residuals (Dunn and Smyth, 1996) were generated using the 'rqgam.check' function of the 'dsm' package (Miller et al., 2015). Randomized quantile residuals transform the residuals to be exactly normally distributed, therefore removing artifacts generated by the link function and making the residuals vs. linear predictor plot easier to interpret.

Sponges (Porifera)

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sponge biomass are presented in Table A1. The R^2 was fair for both models, and slightly higher for the GAM model using the RF-selected variables than for the model using variables correlated at less than 0.7. The deviance explained was higher for the GAM 0.7 Variable model. The AIC/BIC was comparable between the two models. The variable significance for the GAM RF Variable and GAM 0.7 Variables models are shown in Tables A1.2 and A1.3, respectively.

Figure A1.1 shows graphical diagnostics for both models. Both models showed fairly normal residuals and only small patterns in the residuals vs. linear predictor plots. However, the response vs. fitted values plots showed a poor fit between the predicted and actual values for both models. When predicted to the entire extent of the study area, the models showed erroneously high predicted biomass values. High predictions of biomass were not alleviated with the inclusion of latitude and longitude in the model. The predicted surfaces therefore are not presented in this report.

Table A1.1. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sponges in the Newfoundland and Labrador Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.250	0.199
Deviance explained	54.10%	58.90%
AIC	35936.654	35366.241
BIC	36478.898	36454.867

Table A1.2. Results of the GAM RF Variables model built to predict the biomass of sponges in the Newfoundland and Labrador Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Summer Primary Production Average Minimum	8.257	24.480	$< 2 \times 10^{-16}$*
Spring Primary Production Average Minimum	7.465	15.910	$< 2 \times 10^{-16}$*
Bottom Salinity Average Range	7.517	17.730	$< 2 \times 10^{-16}$*
Summer Chlorophyll *a* Mean	8.241	40.620	$< 2 \times 10^{-16}$*
Fall Chlorophyll *a* Minimum	7.227	19.230	$< 2 \times 10^{-16}$*
Bottom Temperature Average Range	7.520	18.800	$< 2 \times 10^{-16}$*
Depth	8.517	40.800	$< 2 \times 10^{-16}$*
Surface Salinity Average Range	8.472	16.740	$< 2 \times 10^{-16}$*

Table A1.3. Results of the GAM 0.7 Variables model built to predict the biomass of sponges in the Newfoundland and Labrador Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Maximum	2.239	10.215	2.130×10^{-6}*
Bottom Temperature Average Maximum	8.433	18.203	$< 2 \times 10^{-16}$*
Annual Chlorophyll a Mean	3.117	8.042	4.450×10^{-6}*
Fall Chlorophyll *a* Maximum	2.863	9.598	7.320×10^{-7}*
Fall Chlorophyll *a* Mean	7.986	12.158	$< 2 \times 10^{-16}$*
Fall Chlorophyll *a* Minimum	7.771	14.304	$< 2 \times 10^{-16}$*
Spring Chlorophyll *a* Maximum	7.059	22.702	$< 2 \times 10^{-16}$*
Spring Chlorophyll *a* Minimum	6.105	11.098	3.800×10^{-14}*
Depth	7.401	43.520	$< 2 \times 10^{-16}$*
Maximum Spring Mixed Layer Depth	8.041	28.962	$< 2 \times 10^{-16}$*
Annual Primary Production Average Minimum	3.179	9.912	5.170×10^{-8}*
Fall Primary Production Average Maximum	7.908	4.455	1.340×10^{-5}*
Fall Primary Production Average Range	4.475	5.964	8.830×10^{-6}*
Spring Primary Production Average Maximum	70.83	13.805	$< 2 \times 10^{-16}$*
Spring Primary Production Average Minimum	3.716	10.823	1.810×10^{-9}*
Spring Primary Production Average Range	7.234	22.490	$< 2 \times 10^{-16}$*
Summer Primary Production Average Maximum	7.794	10.369	3.750×10^{-15}*
Summer Primary Production Average Range	4.235	3.436	3.680×10^{-3}*
Surface Current Average Maximum	7.423	7.016	2.330×10^{-9}*
Surface Salinity Average Range	8.618	22.175	$< 2 \times 10^{-16}$*
Slope	2.219	28.746	4.480×10^{-16}*

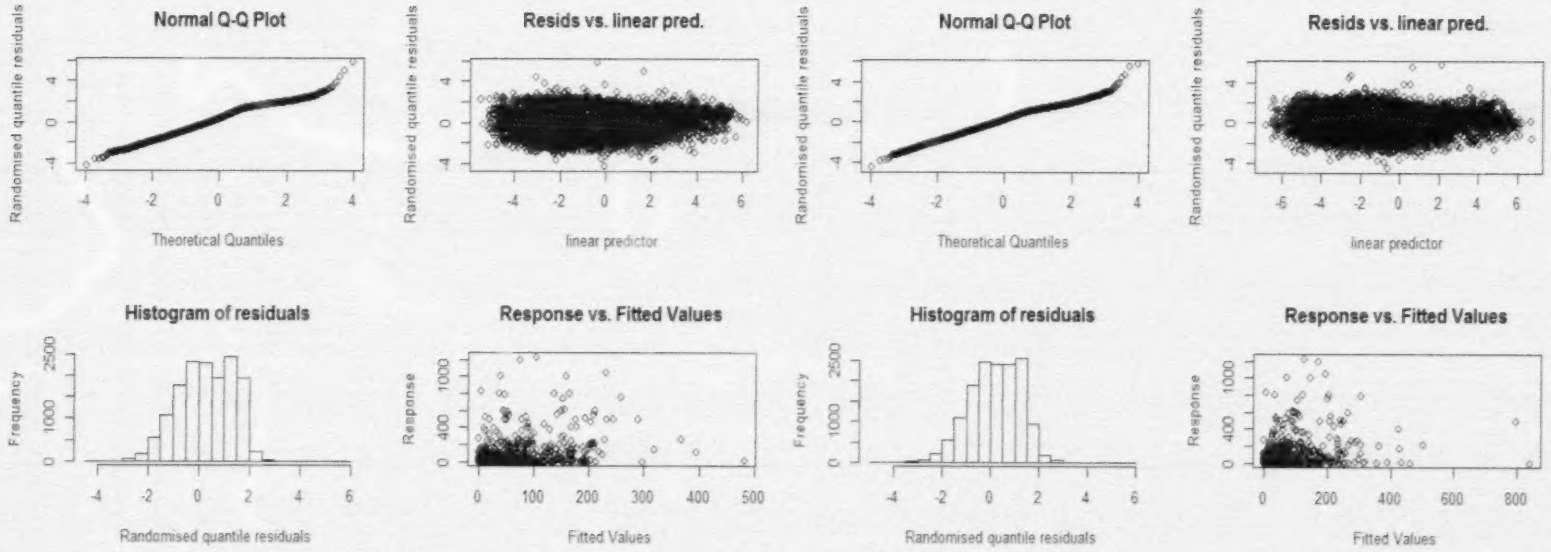

Figure A1.1. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the biomass of sponges in the Newfoundland and Labrador Region.

Sea Pens (Pennatulacea)

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sea pen biomass are presented in Table A1.4. The R^2 and deviance explained were much higher for the model using the variables correlated at less than 0.7 than the model using the RF-selected variables, indicating a good model fit. The AIC/BIC was also lower for the GAM 0.7 Variables model. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.5 and A1.6, respectively.

Figure A1.2 shows the graphical diagnostics for both models. Both models showed fairly normal residuals. The residuals vs. linear predictor plot for the GAM 0.7 Variables model showed patterns indicative of heteroskedasticity. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

The GAM RF Variables model predicted erroneously-high biomass values when applied to the Newfoundland and Labrador study extent, therefore the predicted surface is not presented here. Although performance measures (R^2 and percent deviance explained) were slightly improved, erroneous biomass values were still predicted to occur by a model including latitude and longitude, therefore these results are not considered here. Figure A1.3 shows the predicted biomass surface of sea pens from the GAM 0.7 Variables model. The majority of the study extent was predicted to have low (> 0 – 0.09 kg) sea pen biomass. Higher sea pen biomass was predicted to occur in the Laurentian Channel, and is consistent with the distribution of high sea pen catches (Figure A1.3, right panel). This area was also predicted to have a high biomass by the random forest model (see Figures 29 and 30). The slope south of Grand Banks was also predicted to have a high biomass of sea pens, and is not supported by data observations. The random forest model predicted low-medium biomass in this area.

Table A1.4. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sea pens in the Newfoundland and Labrador Region.

	GAM RF Variables	GAM 0.7 Variables
R^2	0.031	0.308
Deviance explained	45.10%	75.10%
AIC	5089.875	3912.414
BIC	5352.713	4654.459

Table A1.5. Results of the GAM RF Variables model built to predict the biomass of sea pens in the Newfoundland and Labrador Region.The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Maximum Average Winter Mixed Layer Depth	7.282	22.733	$< 2 \times 10^{-16}$*
Fall Primary Production Average Range	3.922	6.238	1.270×10^{-5}*
Surface Temperature Average Range	8.936	80.918	$< 2 \times 10^{-16}$*
Summer Chlorophyll *a* Range	5.260	8.701	2.520×10^{-09}*
Summer Primary Production Average Maximum	8.143	20.152	$< 2 \times 10^{-16}$*

Table A1.6. Results of the GAM 0.7 Variables model built to predict the biomass of sea pens in the Newfoundland and Labrador Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Maximum	1.431	5.634	4.822×10^{-3}*
Bottom Temperature Average Maximum	7.710	23.985	$< 2 \times 10^{-16}$*
Annual Chlorophyll *a* Mean	3.627	13.810	5.780×10^{-12}*
Fall Chlorophyll *a* Maximum	1.614	10.373	2.660×10^{-5}*
Fall Chlorophyll *a* Mean	6.290	11.023	3.090×10^{-14}*
Fall Chlorophyll *a* Minimum	5.325	3.231	3.163×10^{-3}*
Spring Chlorophyll *a* Maximum	2.073	23.137	7.550×10^{-13}*
Spring Chlorophyll *a* Minimum	3.611	4.850	3.740×10^{-4}*
Depth	6.519	64.791	$< 2 \times 10^{-16}$*
Maximum Spring Mixed Layer Depth	4.504×10^{-3}	0.254	0.963
Annual Primary Production Average Minimum	6.649	7.957	3.320×10^{-10}*
Fall Primary Production Average Maximum	3.206	4.045	2.690×10^{-3}*
Fall Primary Production Average Range	1.244	6.058	5.857×10^{-3}*
Spring Primary Production Average Maximum	3.361	6.057	3.940×10^{-5}*
Spring Primary Production Average Minimum	5.284	7.474	1.160×10^{-8}*
Spring Primary Production Average Range	7.275	7.375	5.850×10^{-10}*
Summer Primary Production Average Maximum	6.377	3.439	8.370×10^{-4}*
Summer Primary Production Average Range	5.416	6.032	1.530×10^{-6}*
Surface Current Average Maximum	5.804	4.805	3.260×10^{-5}*
Surface Salinity Average Range	7.454	7.778	1.310×10^{-10}*
Slope	1.395	5.051	8.598×10^{-3}*

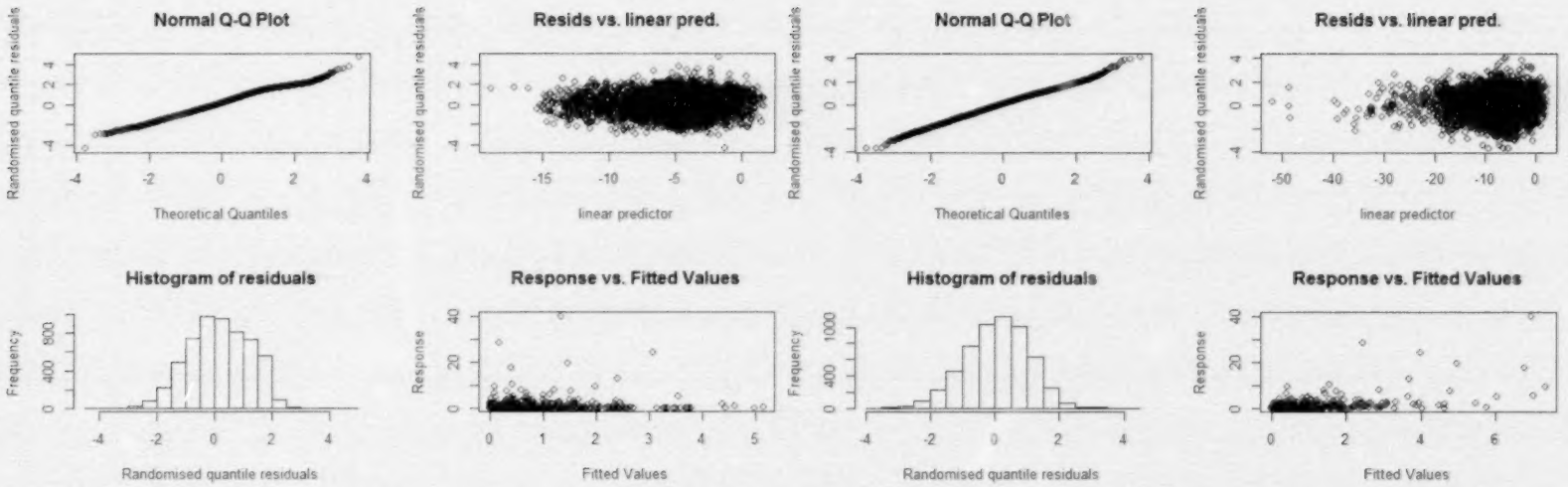

Figure A1.2. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of sea pen biomass in the Newfoundland and Labrador Region.

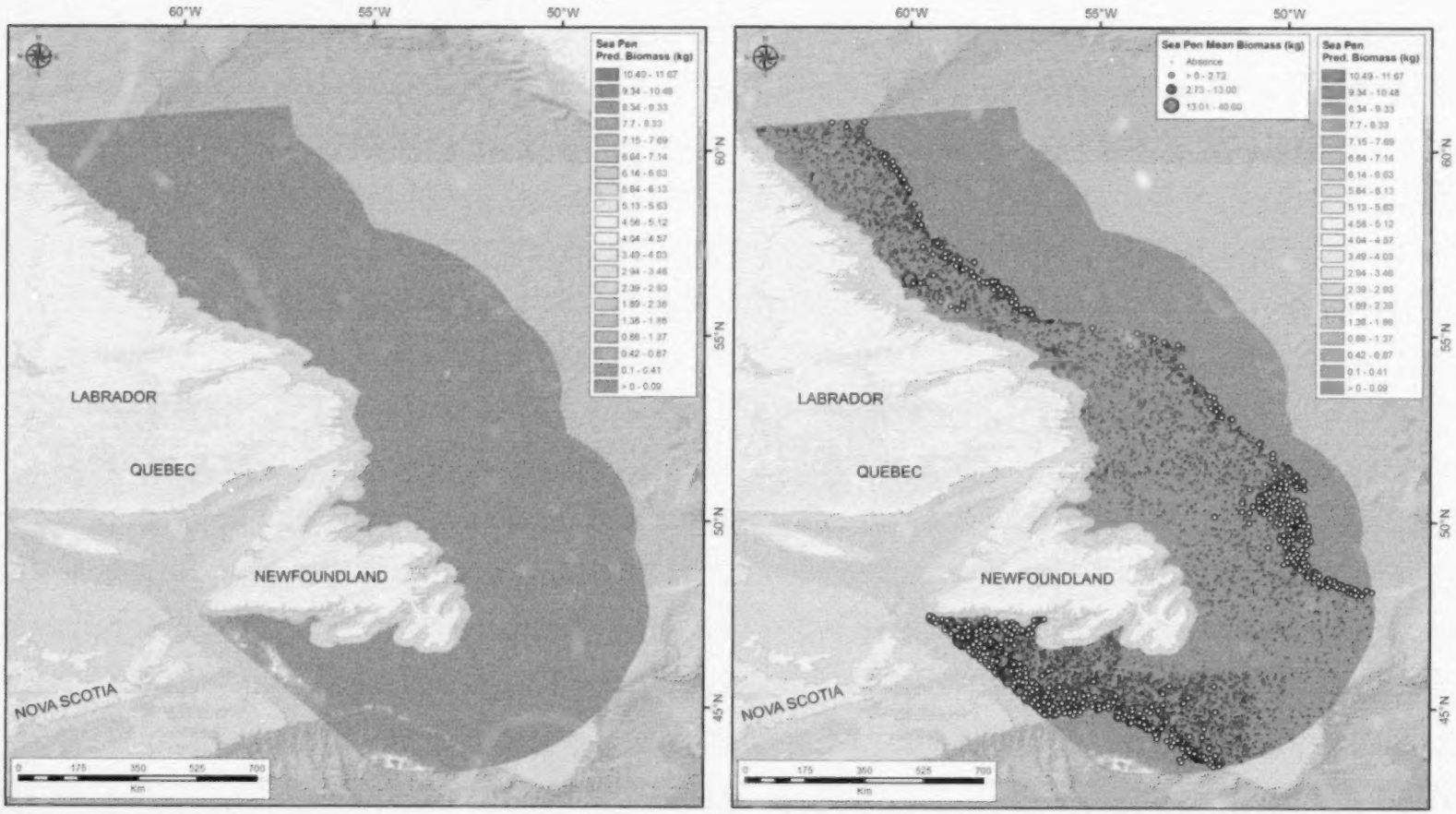

Figure A1.3. Prediction of sea pen biomass (kg) from the GAM 0.7 Variables model in the Newfoundland and Labrador Region. Right map shows the sea pen mean biomass observations overlain.

Large Gorgonian Corals

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean large gorgonian coral biomass are presented in Table A1.7. Both models performed poorly, with R^2 values of 0.040 and -0.154 for the GAM RF Variable and GAM 0.7 Variable models, respectively. Deviance explained was higher for the GAM 0.7 Variable model. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.8 and A1.9, respectively.

Figure A1.4 shows the graphical diagnostics for both models. Both models showed fairly normal residuals and only small patterns in the residuals vs. linear predictor plots. However, the response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

Figures A1.5 and A1.6 show the predicted biomass surface of large gorgonian corals generated from the GAM RF Variables and GAM 0.7 Variables models, respectively. For the GAM RF Variables model, the majority of the study extent was predicted to have low (> 0 – 0.42 kg) large gorgonian biomass. The highest predicted biomass (up to 106.8 kg) occurred on the slope off Saglek Bank. This area of high biomass was associated with a cluster of large biomass values, and is consistent with the random forest results (see Figures 44 and 45). The GAM 0.7 Variables model predicted high biomass of large gorgonian corals in the same area (Figure A1.6). The highest predicted biomass value in this model was 276.78 kg, which is consistent with the maximum mean biomass catch in the raw data (288.97 kg). However, this model poorly predicted the smaller catches that are distributed along the slopes of Labrador and Newfoundland.

Table A1.7. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of large gorgonian corals in the Newfoundland and Labrador Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.040	-0.154
Deviance explained	53%	66.40%
AIC	6405.672	6184.489
BIC	6618.710	6681.857

Table A1.8. Results of the GAM RF Variables model built to predict the biomass of large gorgonian corals in the Newfoundland and Labrador Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Summer Primary Production Average Minimum	5.560	5.780	2.260×10^{-6}*
Fall Chlorophyll *a* Minimum	4.778	9.065	2.660×10^{-9}*
Spring Primary Production Average Maximum	7.645	17.455	$< 2 \times 10^{-16}$*
Bottom Temperature Average Range	7.192	33.587	$< 2 \times 10^{-16}$*

Table A1.9. Results of the GAM 0.7 Variables model built to predict the biomass of large gorgonian corals in the Newfoundland and Labrador Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Maximum	6.906	6.262	7.170×10^{-18}*
Bottom Temperature Average Maximum	6.539	2.762	6.204×10^{-3}*
Annual Chlorophyll *a* Mean	2.509	3.692	0.011*
Fall Chlorophyll *a* Maximum	0.980	3.155	0.065
Fall Chlorophyll *a* Mean	3.212×10^{-4}	0.102	0.994
Fall Chlorophyll *a* Minimum	1.728	1.973	0.133
Spring Chlorophyll *a* Maximum	3.874×10^{-4}	0.038	0.996
Spring Chlorophyll *a* Minimum	1.631	8.897	1.310×10^{-4}*
Depth	5.030	17.942	$< 2 \times 10^{-16}$*
Maximum Spring Mixed Layer Depth	6.022	5.372	5.590×10^{-6}*
Annual Primary Production Average Minimum	4.229	1.852	0.097
Fall Primary Production Average Maximum	1.380	2.752	0.067
Fall Primary Production Average Range	0.815	1.846	0.169
Spring Primary Production Average Maximum	5.729	6.614	1.630×10^{-7}*
Spring Primary Production Average Minimum	2.279	8.071	2.720×10^{-5}*
Spring Primary Production Average Range	4.086	6.014	1.210×10^{-5}*
Summer Primary Production Average Maximum	0.114	0.434	0.765
Summer Primary Production Average Range	2.217	3.274	0.023*
Surface Current Average Maximum	0.696	1.168	0.279
Surface Salinity Average Range	4.560	3.392	0.004*
Slope	1.491	7.590	6.860×10^{-4}*

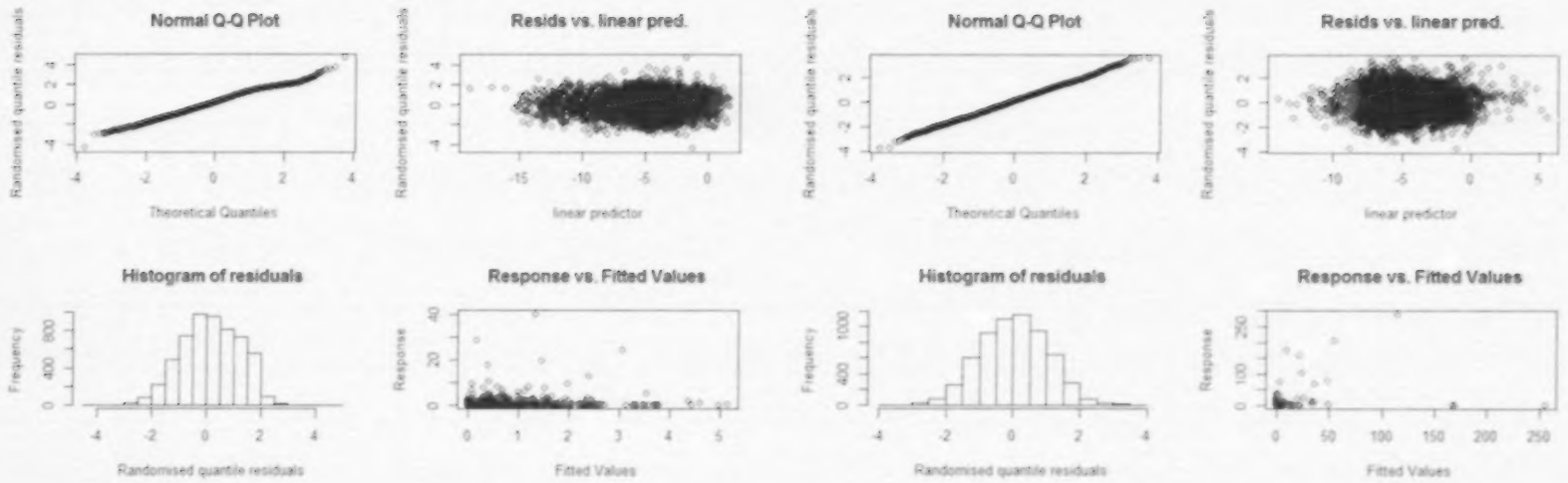

Figure A1.4. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of large gorgonian coral biomass in the Newfoundland and Labrador Region.

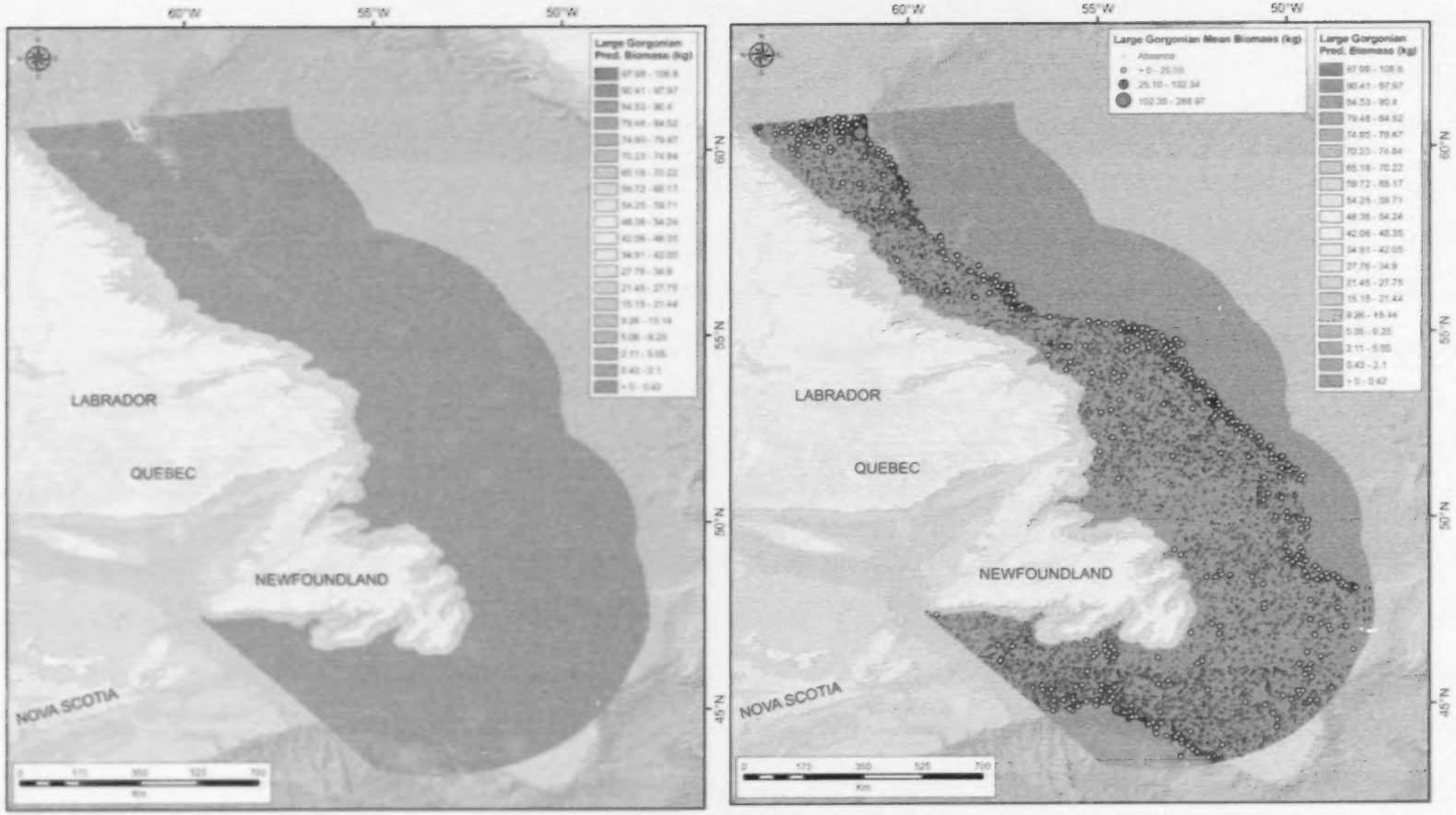

Figure A1.5. Prediction of large gorgonian coral biomass (kg) from the GAM RF Variables model in the Newfoundland and Labrador Region. Right map shows the large gorgonian coral mean biomass observations overlain.

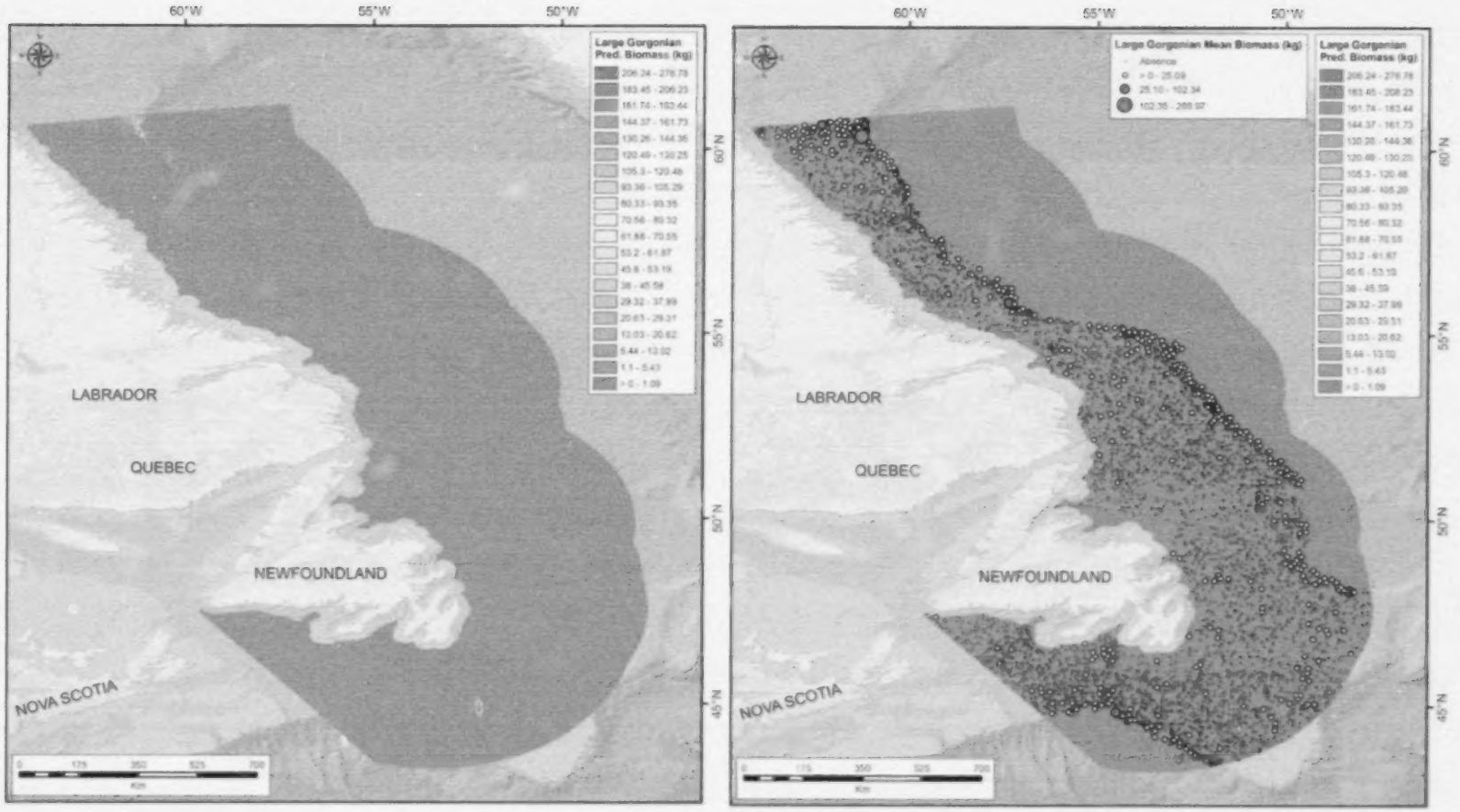

Figure A1.6. Prediction of large gorgonian coral biomass (kg) from the GAM 0.7 Variables model in the Newfoundland and Labrador Region. Right map shows the large gorgonian coral mean biomass observations overlain.

Small Gorgonian Corals

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean small gorgonian coral biomass are presented in Table A1.10. Both models performed poorly, with R^2 values of 0.124 and 0.170 for the GAM RF Variable and GAM 0.7 Variable models, respectively. Deviance explained was higher for the GAM 0.7 Variable model. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.11 and A1.12, respectively.

Figure A1.7 shows the graphical diagnostics for both models. Both models showed normal residuals. The residuals vs. linear predictor plot for the GAM 0.7 Variable model showed patterns in the residuals vs. linear predictor plot indicative of heteroskedasticity. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

Figures A1.8 and A1.9 show the biomass surface of small gorgonian corals generated from the GAM RF Variables and GAM 0.7 Variables models, respectively. For the GAM RF Variables model, the majority of the study extent was predicted to have low (> 0 – 0.42 kg) small gorgonian coral biomass. Higher biomass of small gorgonians was predicted to occur along the slopes of Labrador and Newfoundland. These areas of higher biomass are consistent with the distribution of small gorgonian catches from the RV surveys (Figure A1.8, right panel). The highest biomass of small gorgonians was predicted to occur on the slope southwest of Grand Bank, and is consistent with the results of the random forest model (see Figures 59 and 60). The GAM 0.7 Variables model predicted similar results, however the area of high biomass on the slope southwest of Grand Bank was less intense than that predicted by the GAM RF Variables model.

Table A1.10. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of small gorgonian corals in the Newfoundland and Labrador Region.

	GAM RF Variables	GAM 0.7 Variables
R^2	0.124	0.170
Deviance explained	57%	56.7%
AIC	3610.156	3613.739
BIC	3852.414	3873.916

Table A1.11. Results of the GAM RF Variables model built to predict the biomass of small gorgonian corals in the Newfoundland and Labrador Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Fall Chlorophyll *a* Range	3.509×10^{-5}	0.037	0.999
Surface Temperature Average Minimum	1.978	7.444	2.270×10^{-4}*
Surface Temperature Average Range	8.517	10.861	$< 2 \times 10^{-16}$*
Slope	2.967	16.422	3.960×10^{-11}*
Bottom Salinity Average Minimum	2.716	10.622	1.500×10^{-7}*
Depth	4.954	13.840	1.610×10^{-14}*
Fall Chlorophyll *a* Maximum	3.043×10^{-5}	0.004	0.999
Surface Salinity Average Minimum	4.426	4.433	3.690×10^{-4}*
Fall Primary Production Average Maximum	8.827×10^{-1}	1.480	0.217
Fall Primary Production Average Range	2.608	2.847	0.032*
Surface Temperature Average Maximum	1.372×10^{-3}	0.140	0.986

Table A1.12. Results of the GAM 0.7 Variables model built to predict the biomass of small gorgonian corals in the Newfoundland and Labrador Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Maximum	1.870	1.563	0.201
Bottom Temperature Average Maximum	6.794×10^{-4}	0.293	0.984
Annual Chlorophyll a Mean	1.665	10.623	1.580×10^{-5}*
Fall Chlorophyll *a* Maximum	3.526	3.300	9.170×10^{-3}*
Fall Chlorophyll *a* Mean	2.297×10^{-5}	0.000	1.000
Fall Chlorophyll *a* Minimum	1.423	6.834	1.840×10^{-3}*
Spring Chlorophyll *a* Maximum	9.520×10^{-5}	0.063	0.997
Spring Chlorophyll *a* Minimum	9.313×10^{-1}	3.143	0.068
Depth	3.931	18.118	7.090×10^{-16}*
Maximum Spring Mixed Layer Depth	1.159	2.071	0.132
Annual Primary Production Average Minimum	2.040	8.549	4.940×10^{-5}*
Fall Primary Production Average Maximum	1.946×10^{-5}	0.007	0.999
Fall Primary Production Average Range	3.966×10^{-4}	0.076	0.994
Spring Primary Production Average Maximum	1.434	3.686	0.026*
Spring Primary Production Average Minimum	1.664	4.418	0.010*
Spring Primary Production Average Range	7.551×10^{-5}	0.047	0.998
Summer Primary Production Average Maximum	8.881×10^{-4}	0.053	0.992
Summer Primary Production Average Range	3.324×10^{-5}	0.268	0.996
Surface Current Average Maximum	2.714×10^{-5}	0.090	0.998
Surface Salinity Average Range	7.743	17.521	$< 2 \times 10^{-16}$*
Slope	3.091	18.002	1.800×10^{-12}*

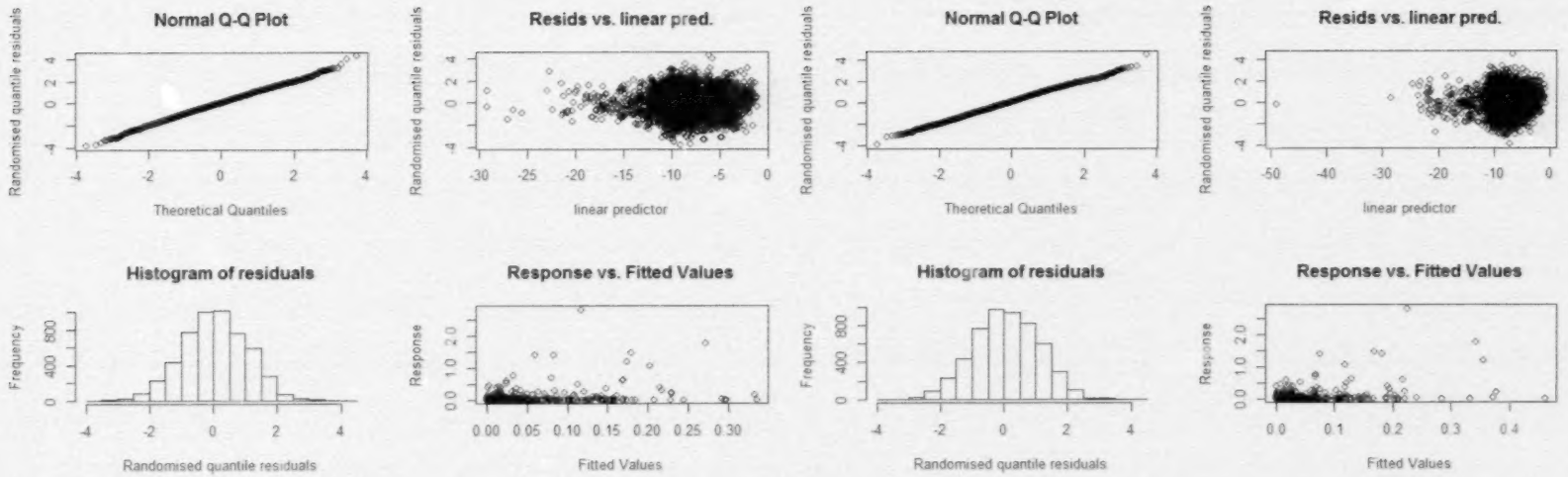

Figure A1.7. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the distribution of small gorgonian coral biomass in the Newfoundland and Labrador Region.

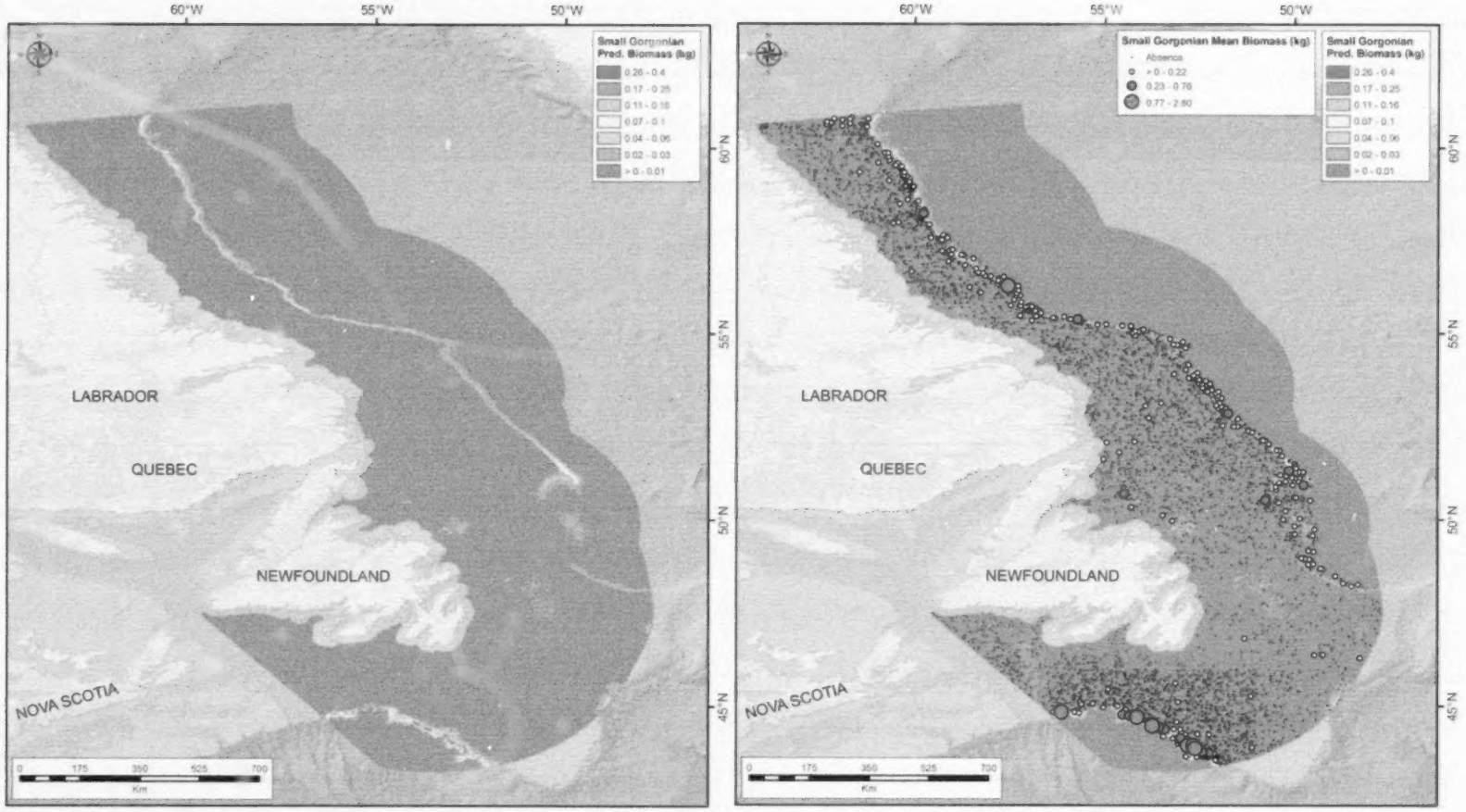

Figure A1.8. Prediction of small gorgonian coral biomass (kg) from the GAM RF Variables model in the Newfoundland and Labrador Region. Right map shows the small gorgonian coral mean biomass observations overlain.

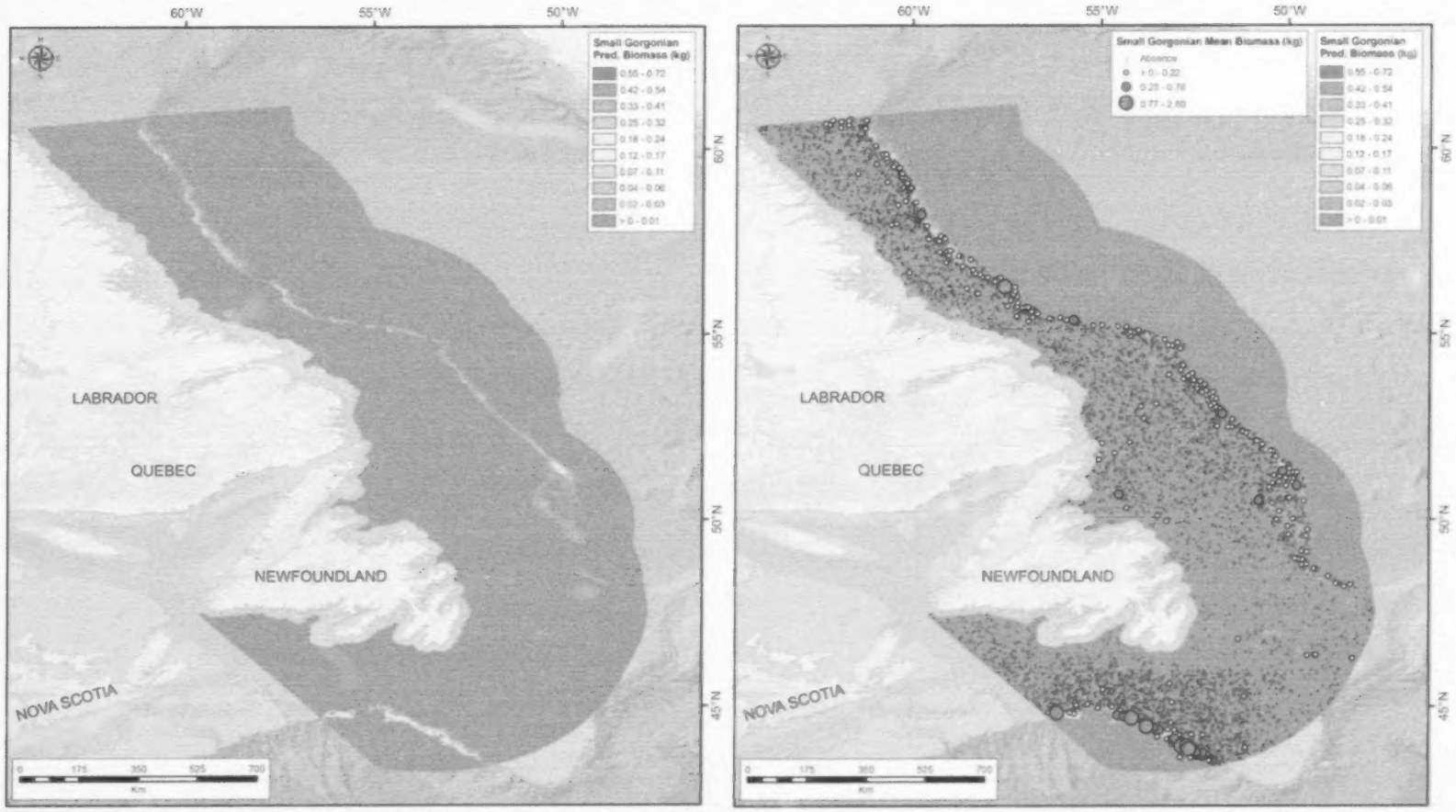

Figure A1.9. Prediction of small gorgonian coral biomass (kg) from the GAM 0.7 Variables model in the Newfoundland and Labrador Region. Right map shows the small gorgonian coral mean biomass observations overlain.

APPENDIX 2

Congruence between Fisheries Observer Data and Species Presence Probability

Fisheries Observer Program Data (FOP) (for more details contact V. Wareham, DFO, NWAFC, St. John's, NL; pers. comm.) from the period of 1996 to 2015 were used to validate the presence probability maps where available. This dataset consisted of 6500 sponge, 1105 sea pen, 592 large gorgonian coral, and 471 small gorgonian coral records. An additional 2406 sponge observer records were obtained from commercial surveys conducted between 1985 and 2001 using shrimp trawl gear, bottom otter trawl gear, longlines, and mid-water trawl gear (S. Fuller, Ecology Action Centre, Halifax, NS; pers. comm.). The overlay of observer data in Newfoundland and Labrador had showed good congruence with the presence probability of sponges (Figures A2.1 and A2.2), sea pens (Figure A2.3), large (Figure A2.4) and small (Figure A2.5) gorgonian corals. For sponges, several FOP records occurred in deep water off the Labrador Slope in an area considered extrapolated and may help to validate the presence probability there. Also, several FOP sponge records from 1985 to 2001 occurring on the Grand Banks of Newfoundland were predicted as absence by the model (Figure A2.2). FOP records for sea pens, and large and small gorgonians were concentrated along the slopes of Newfoundland and Labrador, particularly on the slope off southwest Grand Bank in the 3O Coral Protection Zone. Several large gorgonian coral records were located on the shelf in areas not consistent with prevalence. Sea Pen FOP records were also concentrated on the slope northeast of Newfoundland. This area was in the RF model results and the high presence probability predicted in this area.

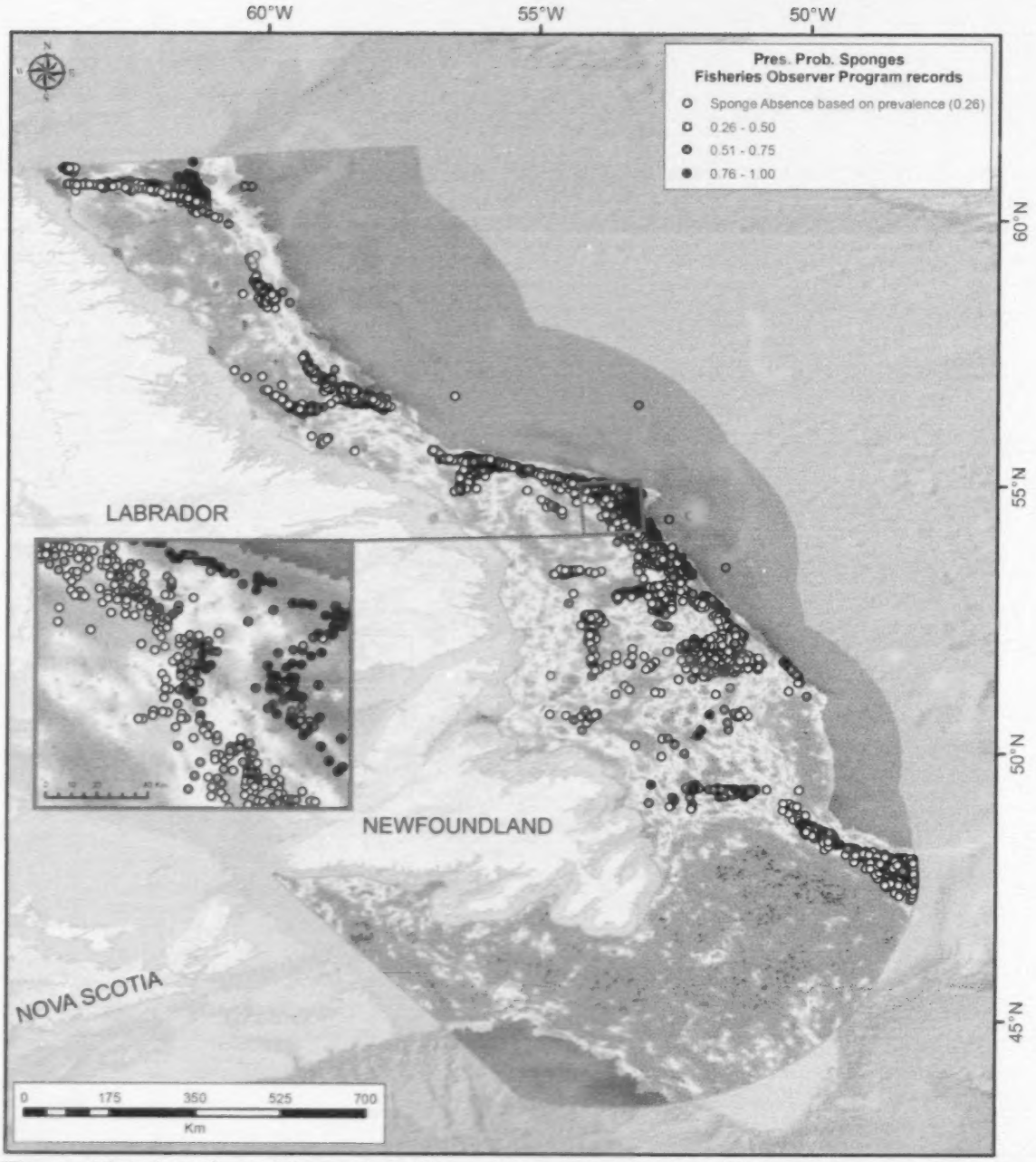

Figure A2.1. Location of the start positions of commercial tows with sponge catches from the Fisheries Observer Program (1996 - 2015) in the Newfoundland and Labrador Region overlain on the sponge RF presence probability map. Also shown are the grey areas of model extrapolation. n = 6500.

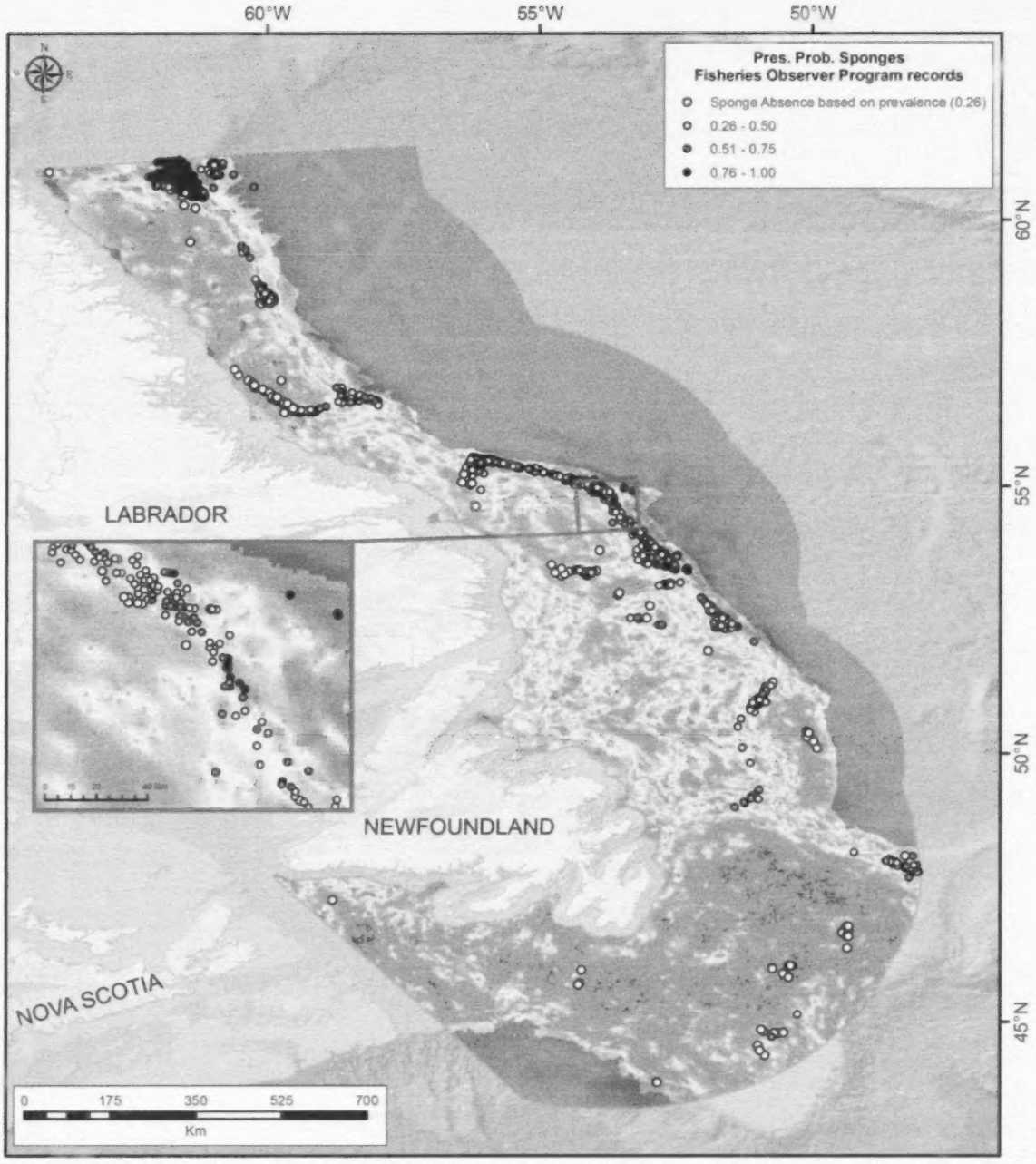

Figure A2.2. Location of the start positions of commercial tows with sponge catches from 1985 - 2001 in the Newfoundland and Labrador Region overlain on the sponge RF presence probability map. Also shown are the grey areas of model extrapolation. n = 2406.

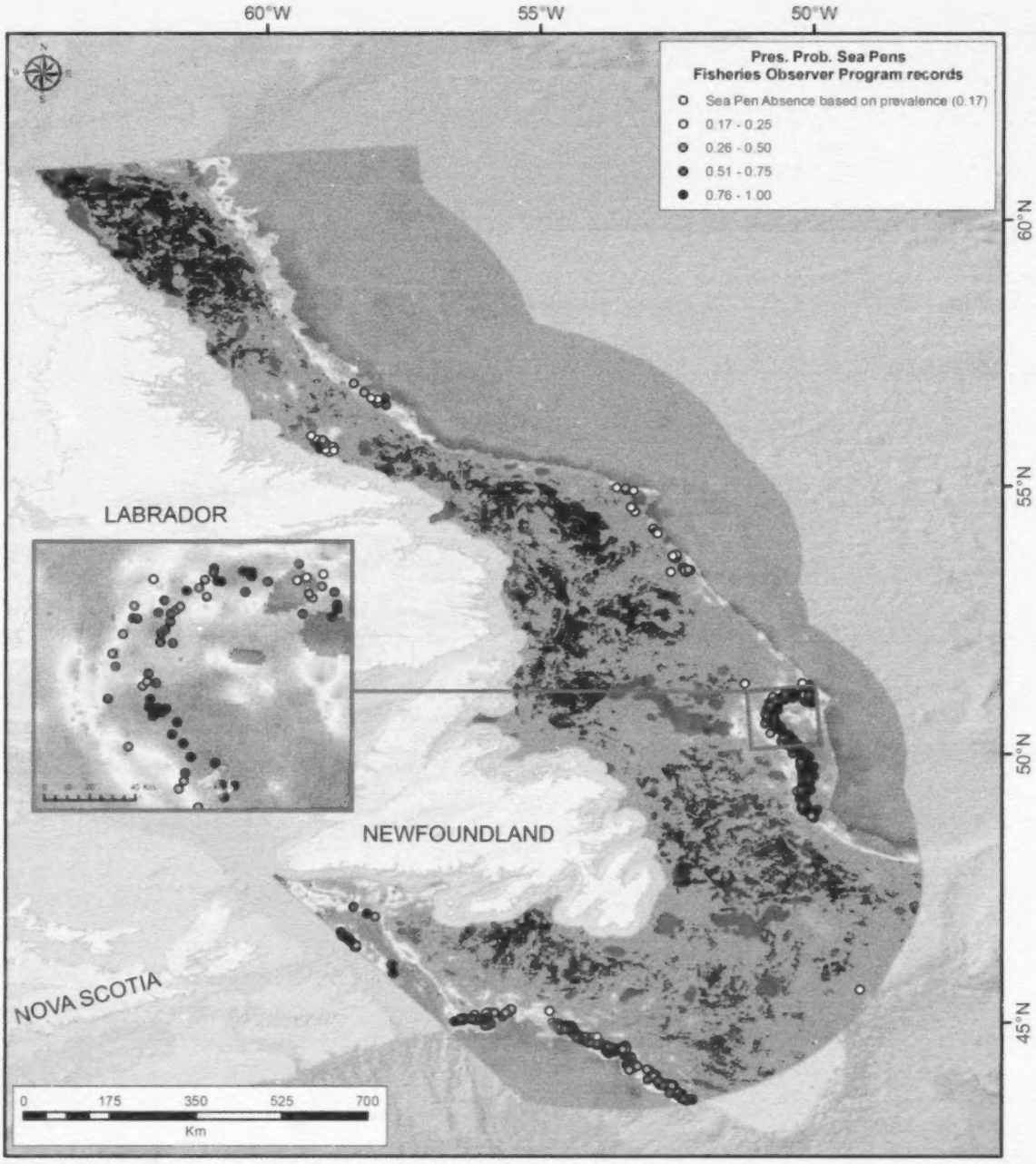

Figure A2.3. Location of the start positions of commercial tows with sea pen catches from the Fisheries Observer Program (2004 - 2013) in the Newfoundland and Labrador Region overlain on the sea pen RF presence probability map. Also shown are the grey areas of model extrapolation. n = 1105.

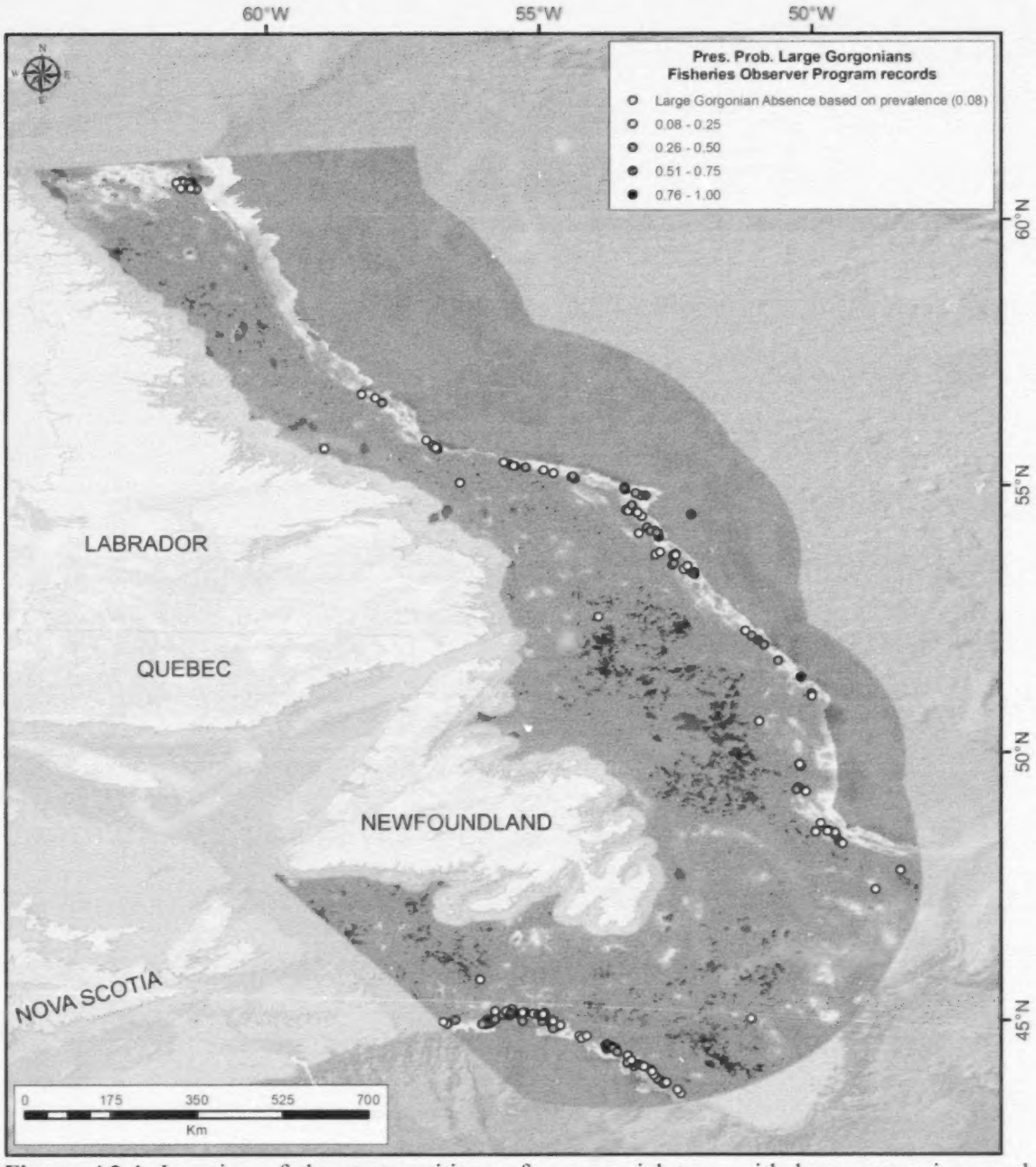

Figure A2.4. Location of the start positions of commercial tows with large gorgonian coral catches from the Fisheries Observer Program (2004 - 2013) in the Newfoundland and Labrador Region overlain on the large gorgonian coral RF presence probability map. Also shown are the grey areas of model extrapolation, which appear dark red or blue when overlain on the presence probability surface. n = 592.

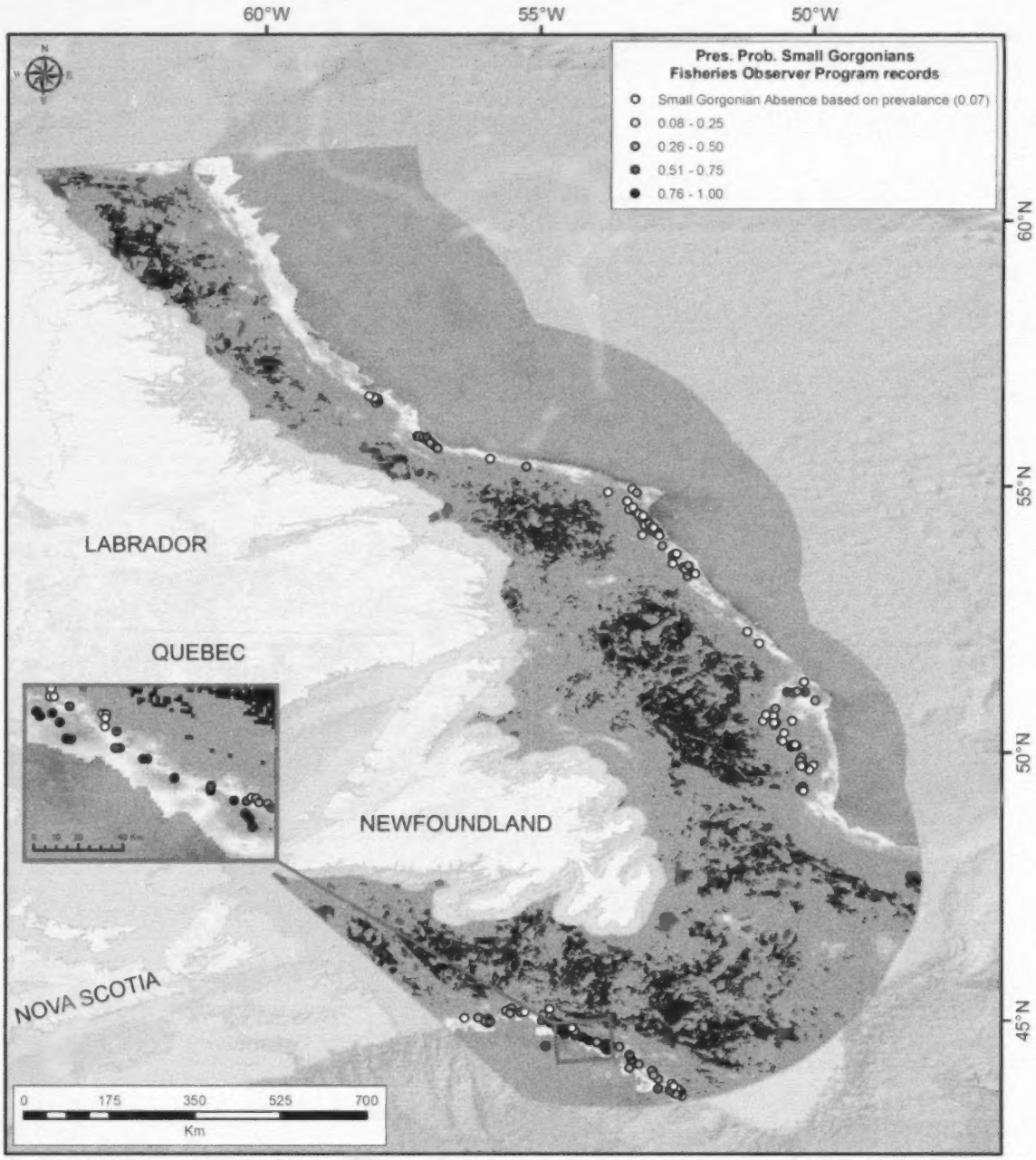

Figure A2.5. Location of the start positions of commercial tows with small gorgonian coral catches from the Fisheries Observer Program (2004 - 2013) in the Newfoundland and Labrador Region overlain on the small gorgonian coral RF presence probability map. Also shown are the grey areas of model extrapolation. n = 471.